EDUCATIONAL
RESEARCH
INSTITUTE OF
BRITISH COLUMBIA

Report No. 82:20

A COMPARISON OF CONVENTIONAL AND RASCH ITEM ANALYSIS APPROACHES APPLIED TO A GRADE FOUR SCIENCE TEST ITEM POOL

John W. Knodel

REPORTS

REPORT NO.: 82:20

TITLE: A COMPARISON OF CONVENTIONAL AND RASCH ITEM ANALYSIS APPROACHES APPLIED TO A GRADE FOUR SCIENCE TEST ITEM POOL

GRANT NO.: DG 281

AUTHOR(S): John W. Knodel

PROJECT DIRECTOR: (if different from author)

PROJECT SITE OR SAMPLE: School District No. 14 (South Okanagan)
School District No. 15 (Penticton)
School District No. 77 (Summerland)

The report is based on a study or project funded by the ERIBC Grants Program. The report may be reproduced in whole or in part, but the author(s) and ERIBC must be credited.

ERIBC takes no responsibility for the work as it was carried out, or the conclusions drawn by the author(s).

This report is available on a loan basis (in B.C. only) or a purchase basis at the cost of $15.30 from the,

EDUCATIONAL RESEARCH INSTITUTE OF BRITISH COLUMBIA
Suite 400 - 515 West 10th Avenue
Vancouver, B.C. V5Z 4A8

Telephone: 873-3801

A COMPARISON OF CONVENTIONAL AND RASCH ITEM ANALYSIS APPROACHES APPLIED TO A GRADE FOUR SCIENCE TEST ITEM POOL

by

JOHN WILLIAM KNODEL

B.Sc., The University of British Columbia, 1968

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

(Educational Psychology)

We accept this thesis as conforming

to the required standard

.....

.....

.....

THE UNIVERSITY OF BRITISH COLUMBIA

September, 1981

©John William Knodel, 1981

ABSTRACT

Research Supervisor: Dr. W. T. Rogers

The purpose of this study was to compare the results of applying conventional and Rasch item analysis approaches to a grade four science test item pool. A 76-item modified version of the pilot tests used to construct the British Columbia Grade 4 Science Assessment Test administered in Spring 1978 was utilized. This item pool was administered to 527 grade four students attending 15 schools located in three adjacent South Okanagan school districts. Eleven booklets were eliminated through application of criteria aimed at controlling for possible effects of speededness. Item analyses were obtained using the 516 remaining booklets.

Preliminary investigations of the test data indicated that it would be best to limit item analyses to the 30-item Concepts and 32-item Processes subtests in the item pool. Coefficient alpha indices and factor analysis data were used to determine the unidimensionality of the subtests. Coefficient alpha indices indicated strong subtest homogeneity. For each of the subtests, however, more than one common factor was found on which there were salient loadings. Study of clusters of items with salient pattern coefficients, however, failed to yield unique definitions of possible different traits being measured by the subtests. It was decided that the subtests were essentially unidimensional and that application of the Rasch model was justified.

The LERTAP computer program was used for conventional item analysis. Four criteria relating to item difficulty, corrected item-subtest point-biserial correlations, distractor-subtest point-biserial correlations, and distractor difficulty were applied.

The BICAL computer program was used for the Rasch item analyses.

Rasch criteria used related to item mean square fit, item discrimination, and item difficulty. For Rasch Approaches I and IV all criteria were used. In Rasch Approaches II and V, the item difficulty criterion was eliminated. Rasch Approaches III and VI used only the mean square fit criterion. For Rasch Approaches I, II, and III, Panchapakesan's correction for guessing formula was used to determine subject membership in the calibration sample. The random guessing level formula was applied in Rasch Approaches IV, V, and VI.

Eight comparisons were made on the subtests resulting from the application of the Conventional Approach and the six Rasch approaches. Four of the comparisons were aimed at the item level. These included the percentage of items rejected by each approach, the efficiency of the Rasch approaches in eliminating items illustrating problems related to conventional criteria, the percentage overlap of rejected items among pairs of different approaches, and the percentage of items rejected solely on the basis of Rasch criteria in the Rasch approaches.

Four comparisons focussed on the subtests as entities. The first involved comparisons of numbers of items in each subtest, subtest means, standard deviations, and score ranges, as well as Hoyt estimates of internal consistency and subtest standard errors of measurement. A second comparison involved correlations of subjects' scores among all Concepts subtests and among all Processes subtests. In a third comparison, correlated t-tests were performed among all possible pairs of Concepts subtests and among all possible pairs of Processes subtests. The final comparison involved the fit of items in the subtests to the Concepts and Processes items used in the final version of the British Columbia Grade 4 Science Test.

The one conventional and six Rasch approaches produced quite different Concepts and Processes subtests as regards specific items selected by each.

Numbers of items in the subtests and consequently subtest characteristics, however, were more related to the stringency of the criteria applied rather than the approach--conventional or Rasch--used to build the subtests.

The number of items in each subtest affected the reliability of the instrument. Use of the Spearman-Brown prophecy formula to adjust the Hoyt internal consistency estimates for the subtests yielded nearly equivalent reliabilities. The content sampling of the subtests was also a function of the stringency of the item analysis criteria used in their construction. Shorter subtests provided a poorer sampling of the content domain than did the longer subtests constructed using more lenient approaches.

While there appears to be a similarity between the conventional p-value and the Rasch item difficulty index, application of Rasch criteria identified items with conventional item-subtest point-biserial problems less effectively. Rasch item analysis proved to be particularly inefficient in identifying items with conventional distractor problems.

Although Rasch approaches produced subtests of equal reliability compared to those built using the conventional approach, and Rasch subtests ordered subjects in essentially the same fashion as subtests built using conventional methods, it was concluded that the conventional approach to item analysis should remain the method of choice. The conventional approach provides information not only to identify poor items but also to improve them. Rasch approaches provide information related to item quality, but do not provide insights for improvement of poor items. Rasch approaches to item analysis should therefore only be applied to large item pools where rejection of items would not seriously affect the resulting instruments' effectiveness in sampling the content domain.

TABLE OF CONTENTS

Abstract..... ii
List of Tables..... viii
List of Figures..... ix
Acknowledgements..... x
CHAPTER I THE PROBLEM..... 1
British Columbia Grade 4 Science Assessment Test..... 3
Test Purpose..... 3
Test Planning and Construction..... 4
Goals and objectives..... 5
Test items..... 7
Pilot test administration..... 8
Item analysis..... 8
Organization of the Present Study..... 10
CHAPTER II PRESENTATION OF BASIC PROCEDURES AND REVIEW OF RELEVANT LITERATURE..... 12
Conventional Item Analysis..... 12
Item Difficulty Index..... 12
Item Discrimination Indices..... 15
Point-biserial correlation coefficient..... 16
Biserial correlation coefficient..... 17
Distractor adequacy..... 18
Choice of item discrimination index..... 19
Correction of item-total correlations for spuriousness..... 22
Test Evaluation..... 25
Internal Consistency..... 25
Hoyt's ANOVA..... 26
Cronbach's coefficient alpha..... 27
Standard Error of Measurement..... 28
The Rasch Logistic Model..... 30
General Latent Trait Theory..... 30
Latent Trait Models..... 30
The Rasch Simple Logistic Model..... 31
Item Characteristic Curves for the Rasch Model..... 34
Assumptions and Robustness of the Rasch Logistic Model..... 34

Adoption of a probability model..... 36
Stochastic independence..... 36
The unidimensionality assumption..... 37
The assumptions of equal item discrimination and lack of quessing..... 40
Estimation of Item Parameters for the Rasch Logistic Model..... 43
Standard error of estimate for item difficulties..... 46
Standard error of estimate for person abilities..... 46
Testing Item Fit to the Rasch Model..... 47
The question of sample size in application of the model..... 50
Test Analysis and Development Using the Rasch Model..... 51
CHAPTER III DESIGN OF THE STUDY..... 55
Item Pool..... 55
The Sample..... 56
Sample Size..... 56
Obtaining the Sample..... 57
Testing..... 59
Data Preparation..... 59
Data Analysis..... 60
Restriction of Analyses to Concepts and Processes Subtests..... 60
Test for Unidimensionality..... 61
Cronbach's coefficient alpha..... 61
Factor analysis..... 61
Item Analysis and Test Construction..... 63
Correction of item-subtest point-biserial correlations..... 65
Elimination of subjects from the Rasch analyses..... 66
Subtest Comparisons..... 67
Percentage of items rejected by each approach..... 67
Percentage overlap of rejected items..... 68
Rasch elimination of items not meeting conventional criteria. 68
Rasch specific items..... 69
Subtest characteristics..... 69
Correlations among subjects' scores on subtests..... 69
Comparison of subtest means..... 70
Determination of fit to the original assessment test..... 70
CHAPTER IV RESULTS..... 72
Final Sample..... 72

Item Analysis and Test Construction..... 73
Subtest Comparisons..... 74
Percentage of items rejected by each approach..... 74
Percentage overlap of rejected items..... 74
Rasch elimination of items not meeting conventional criteria 74
Rasch specific items..... 75
Subtest characteristics..... 78
Correlations among subtests' scores on subtests..... 78
Comparisons of subtest means..... 78
Determination of fit to original assessment test..... 78
CHAPTER V SUMMARY AND CONCLUSIONS..... 84
Summary..... 84
Purpose and Procedure..... 84
Discussion of Results..... 86
Percentage of items rejected by each approach..... 86
Percentage overlap of rejected items..... 87
Rasch elimination of items not meeting conventional criteria 90
Rasch specific items..... 91
Subtest characteristics..... 92
Correlations among subjects' scores on subtests..... 96
Comparisons of subtest means..... 96
Determination of fit to original assessment test..... 97
Conclusions..... 97
Limitations of This Study..... 100
Directions for Future Research..... 100
BIBLIOGRAPHY..... 102
APPENDIX A Test and Administration Materials..... 107
APPENDIX B Correspondence..... 132
APPENDIX C Item Analysis Data..... 137

LIST OF TABLES

1. Selection of Items from the 1977 Grade 4 Science Pilot Tests..... 10
2. Assignment of Items to Goals..... 56
3. Factor Analyses of Subtests..... 63
4. Number of Subjects in Each Item Analysis..... 67
5. Items Used in the Final Version of the Grade 4 Assessment Test..... 71
6. Criteria Used in the Item Analyses..... 73
7. Item Analysis Criteria..... 74
8. Percentage of Items Rejected..... 75
9. Percentage Overlap of Rejected Items--Concepts Subtest..... 76
10. Percentage Overlap of Rejected Items--Processes Subtest..... 76
11. Items Rejected--Conventional Approach..... 76
12. Rasch Approaches--Concepts Subtest: Percentage of Items Rejected According to Conventional Criteria..... 77
13. Rasch Approaches--Processes Subtest: Percentage of Items Rejected According to Conventional Criteria..... 77
14. Percentage of Rasch-Specific Items..... 77
15. Test Characteristics--Concepts Subtest..... 80
16. Test Characteristics--Processes Subtest..... 80
17. Correlations Between Subjects' Scores--Concepts Subtest..... 81
18. Correlations Between Subjects' Scores--Processes Subtest..... 81
19. Correlated t-tests--Concepts Subtest..... 81
20. Correlated t-tests--Processes Subtest..... 82
21. Items Rejected--Concepts Subtests..... 82
22. Items Rejected--Processes Subtest..... 83
23. Percentage of Items in the Final Assessment Test Retained by Each Approach..... 83
24. Percentage Overlap of Rejected Items--Conventional and Rasch Approaches..... 88
25. Percentage Overlap of Rejected Items--Parallel Rasch Approaches..... 88
26. Percentage Overlap of Rejected Items--Rasch Approaches Within and Across Methods of Determining the Calibration Sample..... 89
27. Stepped-up Hoyt Reliability Estimates..... 95
28. Relationships Between Subtest SEM's and $\sqrt{k}$'s..... 96

LIST OF FIGURES

1. Three ICC's Theoretically Obtainable Using the Rasch Model..... 35
2. Raw Score by Item Matrix Illustrating Unidimensionality..... 38
3. Subtest Means (%)..... 94

Acknowledgements

I wish to thank the chairman of my committee, Dr. Todd Rogers, for suggesting the topic for this study and for his support in carrying out the project. Dr. Rogers' courses in statistics and test theory have been a source of challenge and enjoyment throughout my program.

Thanks also to my committee members: Dr. Tom O'Shea, for help in understanding and applying the Rasch model, and Dr. Harold Ratzlaff, whose course in educational research aided me in planning this study, and who made many helpful suggestions during the initial stages of this project.

I would also like to thank Dr. Gaalen Erickson for providing me with important unpublished information related to the British Columbia Grade 4 Science Assessment Test development.

Thanks to the Educational Research Institute of British Columbia which provided financial assistance for this study in the form of a discretionary grant.

Lastly, I wish to acknowledge the important roles played by the superintendents, principals, teachers, and grade four pupils of the South Okanagan (#14), Penticton (#15) and Summerland (#77) school districts, who made the 1979 data collection for the project possible.

CHAPTER I

THE PROBLEM

Test construction has been formalized into several basic steps:

1. Test planning. The content areas and skill categories to be covered in the test are specified according to the purpose the test will serve.

2. Obtaining items. Previously developed items are obtained or new items are written to cover the areas specified in the test plan. More items than will be needed in the final version of the test are assembled so that poor items, as revealed by item-analysis techniques, can be either discarded or improved.

3. Pilot test administration. Items to be analyzed are compiled into a pilot test which is administered to a sample representative of the population for which the final version of the test is intended. The pilot test is administered under conditions which are uniform for all examinees.

4. Item analysis. Test items are analyzed employing an acceptable method of item analysis. Unsatisfactory items are either discarded or improved upon where possible. A revised test is constructed using acceptable or refined items.

5. Revised test evaluation. The revised test is administered to a representative sample of examinees. The test is then analyzed to determine whether it has certain desirable features relating to administration time and reliability. If the revised test meets desired specifications it is accepted as the final version of the test.

Steps 1, 2, 3 and 5 are similar for all types of test construction. There is, however, a choice of methods available for use in Step 4--Item Analysis.

Until recently, the conventional method of item analysis, employing data relating to item difficulty and item/total-test score biserial or point-biserial correlations, has been the standard approach to test construction. High-speed computer programs exist to obtain estimates of item parameters, and considerable expertise has been accumulated in applying this conventional method to test construction problems.

In recent years, there has been much interest in the application of latent trait theory, also referred to as item response theory or item characteristic curve (ICC) theory, to item analysis. For example, the 1977 (Summer) issue of the Journal of Educational Measurement contained six papers on latent trait theory and its applications. Much of the interest in latent trait theory has been fostered by Lord's work begun in the 1950's and the subsequent publication of Lord and Novick's Statistical Theories of Mental Test Scores (1968). This textbook included five chapters (four of them authored by Allen Birnbaum) on latent trait theory.

Until recently, much of the research in latent trait theory was focussed on the one-parameter Rasch logistic model introduced by Georg Rasch (1960; 1966). This model provides data relating to a test item's fit to the latent trait model, item difficulty, and item discrimination. Because of its mathematical tractability, computer programs for application of the Rasch model in test construction were developed early. Consequently, research relating to the Rasch model has been predominant in the literature. This trend is changing today as computer programs for the two- and three- parameter models have now been developed.

Given the dominant use of the conventional and Rasch approaches to item analysis and the apparent lack of information regarding the clear advantage of one approach over the other, the purpose of the present study was to objectively apply conventional and Rasch item analysis procedures to a pool of

test items and to compare the specific nature of items selected/rejected and the overall characteristics of tests resulting from their application. Two series of tests, one constructed on the basis of conventional item analysis, the other series constructed on the basis of item analysis using the Rasch model, were constructed. The same item pool was used to construct the two series of tests. For each of the methods, the respective criteria for item selection/rejection were set according to opinion in current literature and strictly adhered to. Since the conventional approach to item analysis has proven to be useful and is predominant in test construction, the efficiency of Rasch item analysis in identifying items showing problems indicated by conventional item analysis criteria was also determined.

British Columbia Grade 4 Science Assessment Test

The item pool employed in the present study consisted of the items used in the pilot phase of the Grade 4 Science Assessment test administered as part of the general Science Assessment conducted in the Province of British Columbia in the Spring of 1978. Given the centrality of the British Columbia Science Assessment item pool at the Grade 4 level in the present study, a discussion of the development and construction of the Grade 4 Science Assessment test is provided below.

Test Purpose

The purpose of the multiple-choice science assessment tests administered to all grade 4, 8, and 12 students in the Province of British Columbia was to inform educators and the public of strengths and weaknesses in specific areas of science learning, to provide possible insights to educators at the classroom, district, and provincial level as to how to maintain strengths and overcome identified weaknesses, and to provide a data baseline for future

monitoring of student learning in science (British Columbia Ministry of Education, 1979).

The science tests formed one facet of the provincial science assessment program. In addition to achievement tests, instruments to measure students' attitudes towards science, scientists, and science teaching were employed. Questionnaires were used to obtain teachers' opinions about science teaching, the science curriculum, textbooks and supplementary materials, science teacher education, and instructional practices in science. The Science Assessment was part of the total Provincial Learning Assessment Program which includes most subject areas of the public school curriculum in British Columbia. General guidelines outlining procedures for these assessments are presented in Assessment Planning: B.C. Assessment Program (British Columbia Department of Education, 1975).

Test Planning and Construction

The group of educators, or Contract Team, responsible for the science assessment program was charged by the Ministry of Education with the task of gaining specific insights regarding British Columbia students' grasp of science skills considered to be important in the provincial science curriculum. The procedure followed by the Contract Team involved (a) identification of goals and related objectives considered to be important in the science curriculum, (b) construction of tests consisting of items related to the objectives, (c) administration of the tests, (d) scoring and analysis of the test data, and (e) discrepancy analysis. The latter process involved an evaluation of the extent to which students were achieving the objectives held to be important (Rogers, 1979). The Contract Team was aided in their work by a Management Committee, consisting of members representative of the educational community, who provided general advice relating to the assessment and

reviewed the assessment materials produced by the Contract Team. In addition, several ad hoc review committees were empanelled to aid in reviewing and revising assessment materials.

Goals and objectives. Preliminaries to test construction began in the Spring of 1977. Since the focus of the assessment was on specific knowledge and objectives, total test scores were not to be used as measures of science achievement. Such global scores would provide information which was too general to detect specific strengths and weaknesses in science achievement. Therefore, the test was planned to provide meaningful statements regarding the state of learning referenced to several science education goals and related specific objectives.

The Contract Team and Management Committee cooperated in drawing up a framework of goals and related objectives for the series of science tests. Goals were defined as statements of broad, general outcomes of science learning. Objectives were defined as specific statements of behavior which were to be considered as acceptable evidence for having achieved proficiency in a particular goal. Objectives were stated in such a manner that specific test items to measure them could be chosen from existing science tests or could be newly constructed. Level of performance on test items related to objectives subsumed under a particular goal was then to be taken as evidence of mastery of that particular goal.

The test makers were concerned that the objectives chosen for testing should reflect the stated goals of the British Columbia science curriculum. Klopfer's Overall Table of Specifications (Bloom, Hastings, & Madaus, 1971) was chosen as a basis for test construction. The goals and objectives listed in this table were felt to approximate the intentions of the provincial science curriculum. Klopfer's table included the core curriculum goals for science as specified by the Ministry of Education. As well as objectives re-

lated to science knowledge and comprehension, the table also emphasized objectives related to the development of scientific literacy and stressed objectives relating to the process of scientific inquiry.

To better reflect the British Columbia science curriculum, the objectives were incorporated under four overarching goals which were held to reflect current curriculum trends in the province. Some of the objectives were reformulated so that the terminology used was more in keeping with the terms used to describe the British Columbia science curriculum. Both the objectives and the content areas to be covered were rated by the Contract Team in terms of priority for assessment purposes at each of the grade 4, 8, and 12 levels.

The modified table of test specifications was next submitted to the Management Committee for evaluation. On the basis of recommendations from the Management Committee, the Contract Team then proceeded to prepare drafts of tables of specifications for multiple-choice achievement tests at the three grade levels. The specifications table for the grade 4 test called for primary emphasis in testing the process skills of science--observing, classifying, measuring, communicating, and combined processes. A secondary, but still important emphasis in the test was to be on the understanding of scientific facts, concepts, and principles.

The next step in the test planning process was to convene Review Panels of educators and members of the public. The purpose of the Review Panels was to inform the professional and lay public of the purposes and design of the Science Assessment and to obtain feedback from teachers and the public as to what the emphases of the assessment should be. The Review Panels also examined and evaluated the goals and objectives for each grade level which had been formulated by the Contract Team, and evaluated sample items for measuring the objectives. The objectives were rated on a scale from 0 (minimum

priority) to 4 (maximum priority). The median rating of the Review Panels was applied as the final overall rating for each objective.

The above stages in test construction led to some minor modifications in the tables of specifications. The Review Panels had expressed concern that the prototype items relating to the scientific application goal did not seem to relate to practical everyday situations. They also suggested that items dealing with the ability to read and interpret newspaper and magazine articles concerning science-related issues be included under the goal of scientific literacy. The final version of the goals to be included in the science achievement tests was as follows: Goal 1--Understanding Concepts; Goal 2--Skill in the Processes of Science; Goal 3--Application; Goal 4--Safety; Goal 5--Scientific Literacy.

Test items. Since the goals settled upon in the final version of the test table of specifications represented a broad area of learning objectives and the testing time was limited, only a small number of the many possible objectives relating to a particular goal could be tested. The Contract Team had already assembled a bank of multiple-choice questions from established science tests. Where there was an insufficient number of items to test a particular goal or content category, additional items were constructed by the Contract Team. Thus, a more adequate sampling of the content domain relating to a particular goal provided a firmer basis for interpretation of achievement related to that goal. Each item was categorized according to goal, content area, and grade level. The items used came from diverse sources and had varying numbers of response choices. Some items, especially those released from the National Assessment of Educational Progress, included the 'I don't know' option. The Contract Team decided to include this option for all the items used.

Pilot test administration. The item pool for grade 4 initially included 81 items. From this item pool, a two-part pilot test consisting of 37 items in Section A and 38 items in Section B was constructed. During the months of November and December, 1977, Pilot Test A was administered to 172 grade 4 pupils and Pilot Test B was administered to a second sample of 179 pupils. Samples of convenience, selected with consideration of the schools' geographic locations, sizes, and socioeconomic environments were used.

Item analysis. The items in the 1977 pilot tests were subjected to conventional item analysis in order to delete unsuitable items and to improve items for retention in the final version of the test. The items relating to each goal or objective were grouped together and a measure of internal consistency (alpha) was obtained for each item cluster.

Statistical data were also obtained for the individual items in each cluster. Items which were easier than expected were inspected for wording which might have inadvertently given away the correct answer. Correspondingly, items which were more difficult than expected were checked for possible ambiguities which may have confused the more capable students taking the test.

The point-biserial correlation of each option with the total score on each cluster of items was also obtained for each item. The point-biserial correlation of the correct option was corrected for spuriousness. If the correct option of an item believed to measure performance related to a particular goal had a negative point-biserial correlation with the items in that goal cluster, the item became a candidate for omission from the final version of the test. Foils having a positive uncorrected point-biserial correlation also alerted the test builders to possible problems with that item.

The primary consideration in constructing the science assessment tests was content validity. For this reason, items were not necessarily discarded

if the values of their conventional item indices did not fall within "acceptable" limits. In some cases, poor item indices alerted the test developers to problems in the wording of the items which could be corrected. In other cases, items with poor index values were retained if it was felt that they possessed content validity. This is a common practice in test construction, especially in cases where the number of items to measure performance in a particular goal is limited. Thus, empirically obtained statistical information related to an item was given secondary importance to the item's logically judged content validity. To illustrate the role played by human judgement in the test construction, as opposed to strict statistical considerations, it should be pointed out that, even after pilot testing, items were transferred from one goal cluster to another if it was felt that they related more closely to that goal than to the one originally intended. Also, one item which was not pilot tested was included in the final version of the test.

Table 1 provides a summary of the number of items in each goal and objective category of pilot tests A and B. The unbracketed numbers show the number of items retained from each pilot test in the final version of the Grade 4 Science Assessment Test. Bracketed numbers indicate the number of items rejected. Inspection of the technical data resulting from the pilot testing shows that items were included in the final test even though their difficulty level fell outside the range of .20 to .80, or their point-biserial correlations with the appropriate item cluster was less than zero (In pilot test A, three items had a difficulty index greater than .80 and two had a difficulty index less than .20; in pilot test B, seven items had a difficulty index greater than .80; one and two items, in pilot tests A and B, respectively, had negative point-biserial correlation coefficients). In each instance, the items were modified to make them more effective and included in the final version of the test as it was felt they measured important goals and objectives which would otherwise not be measured.

Table 1
Selection of Items from the 1977
Grade 4 Science Pilot Tests

Goal	Pilot Test	
	A	B
Understanding Concepts	13 (2)	13 (2)
Skill in the Processes of Science		
Classifying	5 (1)	--[a]
Observing	--	5
Communicating	6	--
Measuring[b]	5 (2)	1[c]
Combined Skills	--	5 (1)
Application	--	5
Safety	--	5
Scientific Literacy	3	1[c]
Total[d]	32 (5)	35 (3)

[a] Goal clusters having small numbers of items were embedded in one pilot test only.

[b] An additional item, not run in the pilot testing was added to this cluster for the final version of the test.

[c] After pilot testing, some items were moved from one cluster to another for the final version of the test.

[d] The pilot testing provided 67 items for the final version of the test. See note b, above.

Organization of the Present Study

Chapter II discusses the literature relevant to the application of the conventional and Rasch methods of item analysis. Chapter III describes the item pool used to construct the conventional and Rasch tests and the item analysis criteria applied in constructing the different tests. Chapter III also describes the comparisons made between the tests constructed. Chapter

IV discusses the results of the comparisons between the conventional and Rasch tests which were constructed. Chapter V provides a summary of the research and conclusions based upon the findings.

CHAPTER II

PRESENTATION OF BASIC PROCEDURES AND REVIEW OF RELEVANT LITERATURE

The application of conventional test item analysis and the Rasch model in the construction of tests is reviewed in this chapter. Procedures and research relating to each method are outlined. The chapter concludes with a review of the research in which the two methods of test item analysis and their products are compared.

Conventional Item Analysis

Characteristics of individual test items have a direct effect on the characteristics of the total test; all properties of the total test can be directly derived from the properties of the individual items which comprise the test (Magnusson, 1967). The purpose of conventional item analysis in test building is to obtain item parameters, such as item difficulty indices and correlations of items with subtests and total tests, which will be useful in constructing tests having particular desired characteristics. In the case of multiple-choice items, conventional analysis also provides information relating to the effectiveness of the foils. Item analysis is performed in order to exercise some control over the eventual means, variances, reliabilities, and validities of measurement instruments being built (Rentz & Rentz, 1978). In the following sections, the nature of conventional test item parameters and their application in test construction will be elucidated.

Item Difficulty Index

In conventional test item analysis, the basic index of difficulty, or difficulty value, for a dichotomously scored item is the percentage, or

proportion, of subjects passing the item. The proportion, p, is the arithmetic mean of the item. The mean test score for a group of examinees is equal to the sum over items of the item arithmetic means. Consequently, if one wishes to construct a test such that a group of examinees similar to the pilot test group will have a predetermined mean score, items having a mean pilot test sample difficulty equal to the desired mean score divided by the number of items to be included in the test should be selected from the item pool (Lord & Novick, 1968). The predetermined mean score would be approximate only because the item difficulty index of an item is always a reflection of the ability of the group trying the item (Ebel, 1972a).

P-values of individual items are also important in test construction in that they influence the characteristics of test score distributions. The p-values, together with the number of items in the test and the intercorrelations among these items, influence the variance and shape of the distribution of total test scores. Total test variance is the sum of all item variances and item covariances in the test as a whole (Magnusson, 1967, pp. 55-56).

Because item variance, s_i^2, is a function of the frequency of correct response on the individual items, that is, $s_i^2 = p_i q_i$, where $q_i = 1 - p_i$, test variance will be maximized, other things being equal, when the variance of the individual items is maximized. This will occur when $p_i = q_i = .5$, and $s_i^2 = .25$ as a consequence (Magnusson, 1967, p. 26). Thus, the contribution of item variance to total test variance is greatest when an item has a p-value of .5 and approaches zero as the p-value of an item approaches zero or one.

P-values of individual items affect the shape of the distribution of total test scores. If a test consists of many easy items, the shape of the distribution of total test scores for a group of normal ability is likely to be negatively skewed with most examinees being placed in the upper

portion of the distribution of total test scores. If a test contains a disproportionate number of difficult items, the shape of the distribution of total test scores for a group of normal ability will probably be positively skewed with most examinees being placed in the lower portion of the distribution of total test scores. A test developer can correct these extremes by inserting additional difficult items into the former test and easier items into the latter test in order to obtain a more normal distribution of total test scores.

There is no hard and fast rule concerning optimum p-values for items to be included in tests. Some researchers advocate the inclusion of items of middle difficulty in test construction as a means to increase the spread of test scores and to improve test reliability. In one study (Ebel, 1972a), three tests each consisting of 16 items selected from an existing 61-item pool were constructed to have the following characteristics: 1) a test in which the difficulty values of the items were concentrated as close to .5 as possible; 2) a test in which the difficulty values of the items were distributed in a uniform fashion over the entire range of available difficulty values, and 3) a test consisting of extreme difficulty values achieved by including the eight easiest and eight most difficult items. Predictably, the first test was found to have the widest dispersion of test scores and highest reliability of the three tests.

Construction of tests by including only items having a difficulty value near .5 does not, however, guarantee a suitable measurement instrument. Item difficulty is only one of the item statistics to be considered when selecting items for a test. Total test and subtest homogeneity must also be considered in selecting items. Maximum test homogeneity is achieved when items with high intercorrelations are included in a test (Green, Lissitz & Mulaik, 1977). An indicant of high item-intercorrelation is a high item-total test or item-

subtest correlation for each item in the total test or subtest. While p-values of .5 do set an upper limit on correlations of individual items with total test or subtest scores, they do not guarantee high intercorrelations between individual items and total test or subtest scores. Consequently p-values and item-total test or item-subtest correlation coefficients must be considered in tandem when items are being selected for a test (Nunnally, 1967).

Since there is little difference in discriminating power between items having close difficulty values, it is suggested that items with p-values ranging from .5 to .7 (Brown, 1976) or .8 (Stanley & Hopkins, 1972) be included in tests. In usual practice, items with p-values ranging from .2 to .8 are included in tests aimed at groups of examinees in which there is a normal distribution of ability. This was the general procedure followed for the development of achievement tests for the British Columbia Science Assessment (Rogers, 1979).

Item Discrimination Indices

Item discrimination indices provide a measure of how well a test item differentiates between examinees having greater and lesser amounts of ability as regards the trait measured by the item. The criterion used may be the subtest score or the total test score, with the discrimination index being an item-subtest or item-total test correlation coefficient. A high positive correlation is expected between the correct option in a multiple-choice question and the criterion score. Negative correlations are expected for the distractors. This is an indication that subjects getting the correct answer for an item have high scores on the criterion and subjects choosing one of the distractors have low scores on the criterion: students with high ability on the trait being measured get items measuring the trait correct while stu-

dents with low ability on the trait get items measuring the trait wrong.

The point-biserial and biserial correlation coefficients are discrimination indices which can be used with dichotomously scored items. They differ in their assumptions regarding the nature of the trait underlying the dichotomy.

Point-biserial correlation coefficient. The point-biserial correlation coefficient measures the relationship between a dichotomous variable for which there are only two categories of scores and a continuous variable. When the point-biserial correlation coefficient is used, the underlying trait or ability is treated as a true dichotomy: those who pass the item and those who fail the item form two mutually exclusive homogeneous groups. The correlation coefficient, r_{pb}, is a special case of the more general Pearson product-moment correlation coefficient. It can be arrived at by calculating the Pearson product-moment correlation coefficient on the data as it exists, or use can be made of several formulas which have been devised to simplify calculation (Magnusson, 1967, pp. 198-202).

The magnitude of the point-biserial correlation coefficient for an item with a criterion depends on the proportion of subjects passing the item, p, and the proportion of subjects failing the item, $q = 1 - p$. The theoretical maximum value of r_{pb} will tend to diminish as the item-difficulty values diverge from .5. Magnusson points out that, theoretically, the point-biserial correlation coefficient has a magnitude limitation of approximately .8 provided that the continuous criterion variable is normally distributed and the dichotomous variable forms a true dichotomy, split in such a way that $p = q = .5$.

When the assumptions that the underlying continuous variable is normally distributed and the dichotomous variables form a true dichotomy are violated,

the point-biserial correlation coefficient has been found to have values greater than .8. In an investigation involving five different distribution shapes for the continuous variable (Karabinus, 1975), the point-biseral correlation coefficient was found to have values greater than .8 with rectangular distributions of the continuous variable having minimal or no overlap on each of the dichotomized values. Values of r_{pb} greater than .9 were found with peaked bimodal distributions having minimal or no overlap with the dichotomized values. Thus, point-biserial coefficient values greater than .8 can be found when the shape of the continuous variable distribution more closely resembles that of the dichotomized variable.

Biserial correlation coefficient. Biserial correlation coefficients are also used with dichotomously scored items, but use of this coefficient requires the assumption that the dichotomy is an artificial imposition on an underlying ability which in actuality is normally distributed and continuous in nature. There is no justification for use of the biserial correlation coefficient, r_{bis}, with data for which this assumption does not seem reasonable. A basic assumption in measurement theory is that true scores for individuals for a trait have a normal distribution. The splitting of individuals with regard to passing or failing a test item is considered to be an arbitrary process with the locus of the split dependent upon the difficulty of the item. Thus, the biserial correlation coefficient might be more plausible for use with psychometric data than the use of the point-biserial correlation coefficient (Henryssen, 1971).

The biserial correlation coefficient is an estimate of the Pearson product-moment correlation coefficient. Several formulas for calculating r_{bis} from raw score data exist (Magnusson, 1967, pp. 202-206). The biserial correlation coefficient can take on a range of values less than -1 and great-

er than +1. This can occur if the assumption of underlying normality is actually incorrect, or when a small number of subjects is used in testing, producing a platykurtic distribution of the continuously measured variable (Glass & Stanley, 1970). Also, the range or r_{bis} will be less than -1 or greater than +1 if the distribution of the continuously measured variable is leptokurtic.

Because point-biserial correlation coefficients depend upon the p and q values of an item, they tend to diminish as the item-difficulty values diverge from .5. Consequently, the point-biserial correlation coefficient for an item will always be smaller than the corresponding biserial value. The relation between the two types of coefficients is shown in the following equation (Magnusson, 1967, p. 205):

$$\frac{r_{bis}}{r_{pb}} = \frac{\sqrt{pq}}{y},$$

where y is the ordinate of the dividing line between the proportions p and q in a unit normal distribition.

Even under optimum conditions for r_{pb}, when $p = q = .5$, the biserial coefficient value will exceed the point-biserial value by 25% when applied to the same data. As item difficulty values diverge from .5, the discrepancy between biserial and point-biserial values applied to the same item will become even larger.

Distractor adequacy. All distractors, or foils, in a multiple-choice test item should attract responses. The ideal situation would be that all incorrect responses were equally distributed among the incorrect options of an item. Distractors attracting no responses are inefficient deadwood, merely serving to increase the reading time of a test. All distractors should have a negative point-biserial or biserial correlation with the subtest or total test criterion score unless there is an adequate rationale to support

the opposite situation. Thus, effective distractors are chosen by examinees who have a low criterion score. The opposite situation indicates that the item might be miskeyed or ambiguously worded, or that the distractors might be so close in meaning to the correct answer that even examinees with high criterion scores are unable to discriminate among them.

It is important that test item analysis be employed for revision of items as well as for selection of good items. Revision of distractors in a poorly written item can do much to enhance its psychometric properties. In a study involving revised items and newly written items embedded in a test, it was found that all revised items improved in discrimination power (Lange, Lehmann, & Mehrens, 1967). The discrimination index used in the study was the D-index in which the difference between the number of correct and incorrect discriminations is expressed as a percentage of the maximum possible correct discriminations (Findley, 1956). The revised items displayed a higher average discrimination than the new items written for the test. Lange et al, concluded that revision of poor items for subsequent use was more economical than constructing entirely new items for each test administered. They found that it took approximately five times longer to develop a new item than it did to revise a poor item.

Choice of item discrimination index. Controversy regarding the choice of either point-biserial or biserial correlation coefficients in item analysis centers around the fact that the point-biserial correlation coefficient is a function of item difficulty while the biserial correlation coefficient is not. For this reason, it is felt that biserial correlation coefficients will demonstrate less variance between groups of examinees differing in ability. This will not be true of the point-biserial correlation coefficient which necessarily must be low for items found to be either too difficult or

too easy for a given group of examinees. The fact that use of the point-biserial coefficient favours test items of medium difficulty may be advantageous, however, in that extremely difficult and extremely easy items contribute very little to the discriminating power of a measurement instrument (Lord & Novick, 1968).

In one study, nine different types of discrimination index were computed for each of the 120 items equally distributed between two parallel forms of a test (Engelhart, 1965). These discrimination indices included $\underline{r}_{pb}$ and $\underline{r}_{bis}$, both estimated from a computational chart, which required only the upper and lower portions of the score distributions. Intercorrelations among the various discrimination indices used were all high, and for the two forms of the test, correlations between $\underline{r}_{pb}$ and $\underline{r}_{bis}$ were .982 and .987 respectively. Correlations between each index and the first centroid factor of each correlation matrix were high. The correlations between $\underline{r}_{pb}$ and the first factor for each form of the test were .975 and .978 respectively. The corresponding correlations between $\underline{r}_{bis}$ and the first factor were .952 and .960. Applying what were considered to be objective standards for the critical values of the different discrimination indices, it was found that a core of 92 items from the 120-item pool were classified as either satisfactory or unsatisfactory by all indices.

In another study, four language tests, each consisting of two levels for both reading and listening, were administered to 4,300 examinees (Aleamoni & Spencer, 1969). A computer item analysis program was designed to calculate the item difficulty, $\underline{r}_{pb}$, and $\underline{r}_{bis}$ statistics for each item. It was found that the correlations between the $\underline{r}_{pb}$ and $\underline{r}_{bis}$ item statistics were high, with a range of .92 to .99 and a mean correlation of .97. The correlations between each index of discrimination and the difficulty index for an item

were found to be similar with respect to the magnitude, sign, and significance level of the coefficients. The high linear relationship between the two correlation coefficients was taken as an indication that r_{pb} and r_{bis} would generally lead to the same rank-ordering of items in the tests, and the high linear relationship between each of the discrimination indices and the difficulty index did not support the contention that r_{pb} is more highly related to the difficulty level of an item than r_{bis}.

In a more recent study, a 50-item verbal analogy test was administered to 2,311 examinees, from which a random sample of 1,000 subjects was drawn to determine a total of 19 different item discrimination indices, including r_{pb} and r_{bis} (Oosterhof, 1976). The 50 items were ranked separately for each of the indices and the rank-order correlations were computed for each of the possible combinations. The majority of the correlations between indices were greater than .9, with the correlation between the r_{pb} and r_{bis} indices being .934. Two factors accounting for 95.1% of the variance were extracted in a principal factor solution of the intercorrelation matrix. Loadings on the first common factor were considered to be a guage of the degree to which each type of discrimination index related to the factor common to all of the discrimination indices. The first factor loadings for r_{pb} and r_{bis} were .979 and .981 respectively. The first factor accounted for 86.4% of the total variance. A second factor, accounting for less than 10% of the variance was interpreted as being associated with item difficulty. The loadings for r_{pb} and r_{bis} on this second factor were -.207 and .124 respectively. Oosterhof concluded that, since all discrimination indices loaded heavily on one factor, they shared a great deal of similarity. Choice of one index over another, in the light of this research, would minimally affect the products resulting from item analysis. Oosterhof recommended that the test developer select

the index to be used on the basis of computational facilities available and intuitive preference.

Correction of item-total correlations for spuriousness. The point-biserial or biserial correlation of an item with a subtest score or total test score used as a criterion will be spuriously high because, in addition to the item's common factor variance with the whole test, its specific variance contributes to the correlation as well. The magnitude of this artifact of item-test correlations is inversely proportional to the number of items making up the criterion and the size of the intercorrelations between the items. Correction formulas that give the correlation between an item and the total score based on the remaining $n - 1$ items in the criterion have been developed for both point-biserial and biserial correlation coefficients.

Guilford (1953) provided a formula for use in adjusting r_{it}, the point-biserial correlation obtained between item i and the total score with item i included, to obtain r_{ir}, the corrected point-biserial correlation of item i with the remainder of the total score after removal of item i. Guilford pointed out that the amount of correction would be greater when r_{it} was small than when it was large. The smaller item-total correlations are the most problematic when a test builder is deciding what items to include in a test.

While Guilford contended that the 'appropriate' correlation coefficient for r_{it} was the point-biserial coefficient, he stated that the formula could be applied in correcting biserial correlation coefficients once they had been changed into point-biserial estimates by application of a transformation formula.

Henryssen (1963) was concerned with the fact that Guilford's formula did not take into consideration that different items would be correlated with somewhat different sums of the remaining items. These different remainders of items would produce total scores that varied in content, consequently

measuring somewhat different dimensions. In order to correlate the items with the same general factor measured by the total test, Henryssen developed correction formulas for r_{pb} and r_{bis} in which the unique parts of the item variances do not influence the correlation. This was accomplished by substituting average covariances for the item variances.

Henryssen provided an empirical demonstration in which the biserial item-test correlation coefficients were calculated for each item with the remaining $n - 1$ items in a test of n items. Also, corrected biserial item-test correlation coefficients were computed using Guilford's formula and Henryssen's new formula. It was shown that the corrected biserial correlation coefficients obtained using Henryssen's formula were identical to the correlation coefficients arrived at by directly computing each item's r_{bis} with the remaining $n - 1$ items in the test. Corrected correlations obtained by applying Guilford's formula deviated from the directly calculated values.

Cureton (1966) provided a different conceptualization of the problems inherent in correcting item-total correlations for spuriousness. His concern was that the reliability of the remainder of the test would vary inversely with the reliability of the item omitted. To account for this, Cureton derived a formula in which the correlation between an item and a test was arrived at by replacing the removed item with a 'rationally equivalent' item. The corrected item-total correlation is invariant with test variance and reliability. Cureton provides formulas for correcting both r_{pb} and r_{bis}. The formulas should only be applied with factorially homogeneous tests in which the same trait or combination of traits is measured by every item.

In one empirical study (Wolf, 1967), the results of applying Guilford's and Henryssen's formulas in correcting item-total biserial correlations were investigated by using a 69-item science test administered to 171 examinees. Uncorrected item-total correlations, corrected item total correlations obtain-

ed by correlating each item with the remaining n - 1 items in the test, and correlations obtained by applying the Guilford and Henryssen formulas were computed. The test had been randomly assembled and the investigator was thus able to examine the effect of varying the number of test items by establishing samples of 10, 20, 30, 40, 50 and 60 items. Correlations obtained with the Henryssen formula agreed exactly with the correlations obtained by correlating each item with the remaining n - 1 items in the test. Use of the Guilford formula produced some discrepancies but these were not as great as the differences between the uncorrected item-total correlations and the correlations computed between each item and the remaining n - 1 items in the test.

As the test was lengthened, Wolf found that the discrepancies between the uncorrected item-total correlations and the Guilford corrected correlations with the correlations computed between each item and the remaining n - 1 items decreased. By including an identical cluster of ten items in each of the seven tests, the investigator was able to examine the effect of test length on uncorrected item-total biserial correlations as compared to the correlations between each item and the remaining n - 1 items in the test. As test length increased it was found that the differences between the uncorrected biserial correlations and the correlations of each item with the remaining items in tne test diminished for the ten tracer items. It was concluded that corrections for item-total correlations would not be necessary when constructing tests from a 40-item or larger item pool. If corrected correlations were desired, Wolf suggested that the Henryssen formula was the correct one to apply to item-total biserial correlations.

In another study (Berk, 1978), the Guilford, Henryssen and Cureton correction formulas were applied to point-biserial item-total correlations obtained in a test of 20 English mechanics items administered to 124 examinees. Eight additional random samples of 4, 6, 8, 10, 12, 14, 16 and 18 items were

established to study the effects of test length on point-biserial item-total test correlations.

It was found that the Guilford corrected point-biserial correlations were identical to the correlations computed for each item and the remaining $n - 1$ items in the test for all tests. The Henryssen and Cureton correction formulas yielded corrected point-biserial correlations which were greater than the correlations for each item with the remaining items in the test but less than the uncorrected item-total correlation coefficients. In general, the uncorrected, Henryssen corrected, and Cureton corrected correlations deviated less from the correlations for each item with the remaining items in the test as test length increased. There was rapid convergence of Cureton corrected correlations with correlations for each item and the remaining items in a test for test lengths above 10 items. For all test lengths, and for each of the three types of corrected correlations as well as the uncorrected item-total correlations, Spearman rank-order correlations of item ranks with correlations for each item and the remaining items in the test were computed. Use of Guilford's or Cureton's correction formulas yielded item ranks identical to those ranks of correlations between items and the remaining items in a test for tests of all lengths. Use of Henryssen's formula produced different rank-ordering of items in tests of less than 10 items. Berk concluded that correction of item-total point-biserial coefficients of .4 or higher was unnecessary in tests with as few as 18 or 20 items, but that the Guilford or Cureton correction formula should be applied to point-biserial correlations for items with discrimination indices below that point.

Test Evaluation

Internal Consistency

Internal consistency coefficients such as Hoyt's Anova and Cronbach's coefficient α measure the degree to which items in a set are interrelated.

In estimating internal consistency coefficients it is assumed that the items comprising a test are a random sample of items from an item population in which the items measure a single, or unidimensional, trait. High interitem correlations carry with them the implication that a test consists of items which measure a single, or unidimensional trait.

Many researchers feel that internal consistency is best achieved by including items in a test which correlate highly with the subtest or total test criterion. Items can be correlated with the total score for the set of items, and those with low item-subtest or item-total test correlations are culled out. This procedure can be repeated until a highly homogeneous, or internally consistent test is obtained (DuBois, 1970). An alternate approach is to begin with a core of items from an item pool which correlate most highly with the internal criterion and, in successive stages, add other items that correlate highly with the subtest or total test until the desired level of homogeneity is achieved (Nunnally, 1967). To determine whether the task of constructing a homogeneous test has been accomplished, measures of internal consistency such as Hoyt's Anova and Cronbach's coefficient α are then applied.

Hoyt's Anova. Hoyt (1941) outlined a method of obtaining a unique internal consistency reliability coefficient for a test through application of analysis of variance. Any given score on a test, $\underline{Y}_{ij}$, is considered to be a linear combination of the following components: μ, a component common to all persons and to all variables (items); $\pi_{\underline{i}}$, a component associated with the person; $\alpha_{\underline{j}}$, a component associated with the item, and $\varepsilon_{\underline{ij}}$, an error component independent of μ, $\pi_{\underline{i}}$, and $\alpha_{\underline{j}}$, and which is normally distributed with a mean of zero and a variance of σ^2_{ε}. The vector, $\underline{Y}_{ij}$, is thus composed of four independent factors or components:

$$\underline{Y}_{ij} = \mu + \pi_{\underline{i}} + \alpha_{\underline{j}} + \varepsilon_{\underline{ij}} .$$

Assuming that items and examinees are random samples drawn from their respective populations, the two-way random effects model is applied to obtain an ANOVA table with MS_{total}, MS_{items}, $MS_{persons}$, and MS_{error} terms. MS_{error} is the mean square associated with the persons-by-items interaction, $MS_{persons \times items}$. The reliability coefficient is obtained by dividing the difference between $MS_{persons}$ and MS_{error} by $MS_{persons}$ (Stanley, 1971, p. 399).

Hoyt's formula for internal consistency reliability gives the ratio of true variance for persons, that is, variance not due to the unreliability of the measurement instrument, to total test variance. The standard error of measurement is obtained by taking the square root of the quotient of SS_{error} divided by $k - 1$, where k is the number of persons taking the test. Hoyt pointed out that his method of estimating test reliability yielded identical results to those arrived at with application of the Kuder-Richardson formula 20, KR 20, (Kuder & Richardson, 1937) if items were scored dichotomously--1 if correct and 0 if incorrect.

Cronbach's coefficient alpha. Coefficient alpha is also a two-way analysis of variance approach to obtaining a coefficient of internal consistency for a test (Cronbach, 1970). When applied to the same dichotomously scored test taken as a whole, internal consistency coefficients obtained using KR 20, Hoyt's Anova, and Cronbach's coefficient alpha yield equivalent results. Cronbach's alpha is a more general formula, however, in that it can be applied to estimate an internal consistency coefficient for a test which is a composite of subtests (Cronbach, 1951). Tests divisible into distinct subtests are broken up into these before application of the formula. To obtain interpretable subtest scores, the best set of items is one having a large coefficient alpha and is not further divisible into smaller clusters of items. The alpha coefficient is an indication of how well the subtests in a total test tend to measure the same factor.

The Hoyt index is computed over subtests, while the alpha coefficient is an index of internal consistency for subtests. The two indices will only be equivalent when the covariances of subtest items, computed within each subtest, all equal zero. Subtest items correlating positively with each other in each subtest will yield a Hoyt value greater than alpha. Alpha will be greater than the Hoyt index if subtest items correlate negatively within each subtest. Since, in designating subtests, one wishes to include items having high intercorrelations, the former situation will normally arise. Coefficient alpha represents a lower bound as an index of common-factor concentration in a test, and an upper bound as an index of first-factor concentration among the items (Cronbach, 1951).

Standard Error of Measurement

In conventional test theory it is hoped to approximate a person's true score on the trait being measured by a test. Because a person's test score is a sample score, it will deviate from his/her true score because of differences in the items used in the parallel forms of tests and because of such measurement errors caused by differences in examiner/examinee interactions, changes in examinee health and motivation, clarity of instructions, outside disturbances, guessing, and subjectivity in scoring (Stanley, 1971, p. 364). The measure of this deviation from a person's true score is called the standard error of measurement (SEM) of the test (Magnusson, 1967, p. 79).

In conventional, or classical, test theory the SEM is an estimate of the average standard deviation of the distribution of measurement errors made when we allow a given individual's obtained score on each one of a theoretically infinite number of parallel tests to represent his/her true score. Measurement errors are assumed to be normally distributed, mutually independent, and independent of the true score which they represent. The last assump-

tion is particularly debatable and has let to the development of different equations for the SEM in which the assumption of equal variance of error distributions for varying true scores is not made (Magnusson, 1967, pp. 82-86).

The standard error of measurement for a test, as defined above, is directly proportional to the square root of the number of items on the test. This relationship has been proved mathematically (Gardner, 1970) and has been substantiated through empirical investigation (Lord, 1959). Lord investigated the relationships between the standard error of measurements for 50 tests of moderate difficulty and test means, standard deviations, item difficulties, and number of items. He found that the number of items in a test was the only significant factor relating to the test standard errors of measurement. The correlations between test standard errors of measurement, when reliability of the tests was determined using the KR 20 formula, and the square root of the number of items on the tests was .996. The best fitting regression line yielded the relationship SEM = $.432\sqrt{n}$, where n is the number of items on a test.

In an empirical study to test Lord's findings, Cureton, Cook, Fischer, Laser, Rockwell, & Simmons (1973) investigated the relationship between the SEM and the square root of the number of items in a test for six different tests. These researchers found that the regression coefficient, when Lord's earlier data was merged with their data, was .417, a somewhat lower value than Lord's constant of proportionality.

Ebel (1972b) provides a rational explanation of why increasing the number of items in a test should improve test reliability:

1. The true component of a score is proportional to the number of equivalent items that contribute to it.

2. The error component of a score is proportional to the square root of the number of equivalent elements that contribute to it.

Magnusson (1967) describes the same principle in somewhat different terms:

> While the true variance increases as the square of the number of times the test is increased in length, the increase in error variance is directly proportional to the increase in test length. This means that, when the test increases in length, the true variance represents a greater share of the total variance. This in turn means that the test will become more reliable. (p. 72)

The Rasch Logistic Model

General Latent Trait Theory

Latent trait theory, also referred to as item response theory, or item characteristic curve (ICC) theory, assumes that examinee performance in testing situations is predictable or explainable by defining examinee characteristics, termed traits, and using text scores to predict or explain performance on tests. The traits or abilities are not directly measurable but are inferred from test performance (Hambleton, Swaminathan, Cook, Eignor & Gifford, 1978). Latent trait models are mathematical models which attempt to explain, as a concise mathematical function, the relationship between examinee performance on a test and the unobservable latent traits or abilities which underlie this performance (Hambleton & Cook, 1977).

When a single latent trait is assumed to underlie performance on a test, the mathematical function that relates the probability of success on an item to the underlying trait measured by the item set or test containing the item is called an item characteristic curve, or trace line. The item characteristic curve is the nonlinear regression function of item score on the latent trait underlying the test (Hambleton & Cook, 1977). The ICC graphically depicts the relationship between an examinee's position on the latent trait measured by a test and the probability of passing an item in that test.

Latent Trait Models

Three latent trait models have been developed. Each model is an attempt to mathematically specify, as concisely as possible, what occurs when a per-

son possessing a given amount of ability related to a particular trait attempts a test item believed to measure that trait. The models differ in the number of variables considered necessary to provide an accurate description of the outcome of the encounter between examinee and test item. All three models employ only one person parameter, or ability parameter, in their formulation of the ICC function. They differ in their number of item parameters considered to be necessary in defining the function. The most complex model includes three item parameters--item difficulty, item discrimination, and a parameter for guessing. A less general two-parameter model omits the guessing parameter in predicting an examinee's success on an item. The Rasch model is the simplest of the three conceptualizations, employing only the item difficulty parameter in its mathematical description of the ICC.

Until very recently, the Rasch model has been the most heavily investigated of the three models. The other two models require complex scoring rules and mathematically unwieldy solutions are necessary to obtain values for the additional item parameters. Currently, however, rapid and efficient computer programs have been developed for the two- and three-parameter models as well.

The Rasch Simple Logistic Model

The Rasch model was developed independently of the other latent trait models by the Danish mathematician Georg Rasch in the early 1960's. In formulating his model, Rasch made three assumptions (Rasch, 1966):

(a) To each situation in which a subject ($s = 1, 2, \ldots n$) has to answer an item ($i = 1, 2, \ldots m$) there is a corresponding probability of a correct answer ($X_{si} = 1$) which we shall write in the form

$$Pr(X_{si} = 1) = \frac{R_{si}}{1 + R_{si}}, \quad (R_{si} \geq 0).$$

(b) The situation parameter, R_{si}, is the product of two factors,

$$R_{si} = A_s W_i,$$

where A_s pertains to the subject and W_i to the item.

(c) Given the values of the parameters, all answers are stochastically independent. (p. 50)

A_s, the subject parameter, may take on any value greater than zero and is a measure of the ability of the subject with respect to the kinds of items being answered. Higher A_s values indicate greater ability. W_i, the item parameter, may also take any value greater than zero and measures the simplicity, or easiness, of an item. Its reciprocal, $D_i = 1/W_i$, measures the difficulty of an item, with higher values indicating items of greater difficulty. If D_i is used instead of W_i, the Rasch equation becomes:

$$Pr(X_{si} = 1) = \frac{R_{si}}{1 + R_{si}} = \frac{A_s W_i}{1 + A_s W_i} = \frac{A_s / D_i}{1 + A_s / D_i} = \frac{A_s}{A_s + D_i}.$$

A_s and D_i are measureable on an underlying ratio scale ranging from 0 to $+\infty$.

The simple logistic form of the Rasch equation may also be derived from the above relationships (O'Shea, 1979). This equation is as follows:

$$Pr(X_{si} = 1) = \frac{e^{a_s - d_i}}{1 + e^{a_s - d_i}},$$

where a_s is the natural logarithm of the person, or subject, ability parameter, and d_i is the natural logarithm of the item difficulty parameter.

The above formulation of the Rasch simple logistic model is common in the current literature (Hambleton et al, 1978). It is the difference between the person ability parameter, a_s, and the item difficulty parameter, d_i, which affects the probability of a successful outcome when a person of

given ability attempts an item of given difficulty. The parameters, a_s and d_i, can theoretically vary from $-\infty$ to $+\infty$ and consequently so can their difference (Wright, 1977a). The unit of measurement for both parameters is the 'logit', which in practical applications usually ranges from -3 to +3 (Rentz, 1979). Negative values of a_s and d_i indicate persons of lower ability and easier items, respectively. Probability of a correct answer to an item, as determined by the logistic equation, remains in its restricted range between zero and one, because the difference between the parameters is applied as the exponent to e, the base of the natural logarithms.

The following examples illustrate how the interplay between a_s and d_i values affect the probability of a correct response to an item (Wright, 1977a). Suppose a person of high ability, $a_s = 3$, attempts an item of average difficulty, $d_i = 0$. The log odds for success on this item is equal to $a_s - d_i = 3 - 0 = 3$, and the value of the probability expression, $Pr(X_{si} = 1)$ is .95. Now suppose a person of average ability, $a_s = 0$, attempts a very difficult item, $d_i = 3$. The log odds for success on the item for that person will be $0 - 3 = -3$, and the probability is .05. If the ability of the person answering the item matches the difficulty of the item exactly, the log odds for success on the item is equal to 0 and the probability for success on the item is .5.

The probability expression illustrates that the greater the positive differences between a_s and d_i becomes, the greater a person's chance for success on the item. The greater the negative difference becomes, the lower is the person's chance for success. When person ability and item difficulty are matched perfectly, there is a 50% chance for success on the item.

It can be demonstrated that Rasch person ability parameters and item difficulty parameters are measured on a ratio scale. Suppose three persons, A, B and C, with abilities 1, 2 and 3 respectively, attempt the same item

which has a difficulty of .5. The odds on success for the same item would be $e^{0.5}$, $e^{1.5}$ and $e^{2.5}$ for persons A, B and C, respectively. The values are 1.65, 4.48 and 12.18. In each case, the odds on success for a given person on the item compared to the odds on success for the person of next highest ability results in the ratio 4.48/1.65 = 12.18/4.48 = 2.72. Thus, persons differing by equal intervals on the Rasch ability scale display the same ratios for odds on success with an item of given Rasch difficulty. Similar examples, in which the ability of a given person is held constant and that person's odds on success with items having Rasch difficulty estimates differing by equal intervals could be calculated. The Rasch model is the only latent trait model which permits this sort of interpretation of person ability and item difficulty parameters (Hambleton et al, 1978).

Item Characteristic Curves for the Rasch Model

Figure 1 depicts three item characteristic curves which are compatible with the Rasch model. The values of the ability parameter, a_s, are plotted along the abscissa of the graph in the practical range, -3 to +3. The ordinate shows the probability of passing an item of given difficulty as it relates to increasing ability. The ICC's are smooth and monotonically increase as ability increases. They have equal slopes because the Rasch model assumes equal discriminating power for all items. In ICC theory, the item difficulty index is defined as the ability value required to have a 50% chance of obtaining the particular item in question correct. The difficulty index of the three items shown is -1, 0 and +1, respectively.

Assumptions and Robustness of the Rasch Logistic Model

Assumptions are the necessary conditions for the consequences of a proposed model to be true (Rentz, 1979). Rasch's three original assumptions--a) adoption of a probability model; b) definition of the situation parameter,

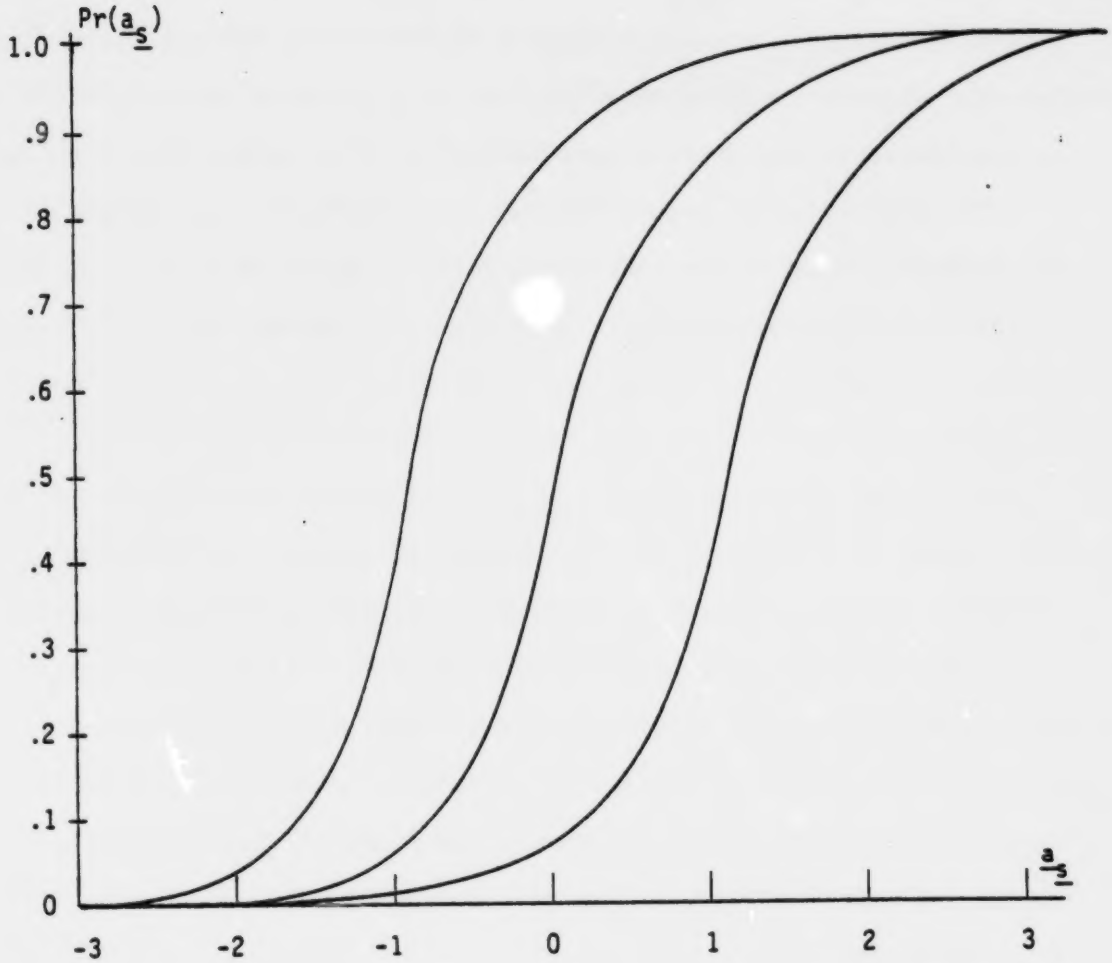


Figure 1. Three ICC's theoretically obtainable using the Rasch model.

R_{si}, as being the product of two factors only, person ability, A_s, and item difficulty, W_i, and c) stochastic independence of answers, were the only assumptions made by him in developing the model (Rasch, 1966). In addition to the original assumptions made by Rasch, the assumptions of unidimensionality of the trait being measured, lack of guessing on multiple-choice items, and equal item-discriminations have arisen from Rasch's second assumption. The latter triad of assumptions has been involved in most of the empirical investigations concerning the robustness of the Rasch model. The robustness of a model is the condition that exists when the consequences of a model hold although all of the assumptions of the model may not have been strictly met (Rentz, 1979).

Adoption of a probability model. In Rasch's first assumption, a probability, or stochastic, model is adopted as a convenient device to describe the interaction between an examinee and a test item. Rasch presented a model which described the probability of obtaining a test item correct as a function of the item's difficulty and the examinee's ability with respect to the trait being measured by the item. The model appears to be compatible with reality. Rasch, however, does not take a position on the question of whether or not test item responses can be ultimately explained in causal terms (Rasch, 1966).

Stochastic independence. In applications of the Rasch model to data, all answers must be stochastically independent, or exhibit 'local independence'. This means that all items must meet a criterion of inter-item independence in that the response given by a person to one item does not affect the response given by that person on other items in the test (Whitely & Dawis, 1974). Thus, multi-part, or chained test items are not compatible with the application of the Rasch model. Also, tests analyzed using the Rasch procedure must essentially be power tests uninfluenced by speededness. Rasch (1966)

encountered the problem of speededness when his model failed to fit the data from tests in which examinees were working under a time stress. When he stratified the examinees according to overall working speed, the data fit the model. Thus, item scores must be statistically related to each other only through the underlying trait measured by a test. When items are not attempted on a speed test, performance becomes a function of an extraneous factor other than the latent trait being measured.

Another facet of stochastic independence is inter-person independence (O'Shea, 1979) which implies that there must be no copying or cheating existing under group testing conditions. The probability that two examinees get a particular test item correct must relate only to their abilities on the latent trait being measured and cannot be influenced by any other factors.

The unidimensionality assumption. In the application of latent trait models, it is typical to assume that the latent space is unidimensional--items in a test are homogeneous, measuring only a single ability or trait (Hambleton & Cook, 1977). Rasch (1966) did not explicitly state this assumption but it is a necessary assumption in the application of latent trait models in general (Lord & Novick, 1968). Whitely and Dawis (1974) present a definition of unidimensionality as they believe it applies to the Rasch model. Consider:

> ... an item by score group matrix in which k items are ordered by their difficulty level and $k - 1$ score groups by obtained raw scores. The score groups for which all items are either passed or failed are excluded from the matrix, since these extreme score groups provide no differential information about the items. The cell entries represent the probability, P_{ij}, that item i will be passed by the score group j. ... unidimensionality implies that for each item, P_{i1} is less than P_{i2} and P_{i2} is less than P_{i3} and so on to $P_{i,k-1}$, so that the probability of passing the item increases regularly with the total score. Each item, then orders the subjects in the same way.
>
> Another implication required for conjoint measurement of subjects and items, is that items are ordered the same way within each score group. On the item by score group matrix, this im-

plies that P_{1j} is less than P_{2j} and P_{2j} is less than P_{3j}, etc. to P_{kj} within each score group. (pp. 164 - 166)

The above conceptualization of unidimensionality can be depicted diagrammatically as shown in Figure 2 (O'Shea, 1979). The diagram shows the matrix of probabilities for the cells of a test of k items, the items having been arranged in order of increasing difficulty. For the condition of unidimensionality to be satisfied, the items must be ordered the same way within each score group as follows,

$$P_{i1} > P_{i2} > P_{i3} > \cdots > P_{ik}.$$

Also, each item must order subjects by membership in a score group in the same way as follows,

$$P_{1j} < P_{2j} < P_{3j} < \cdots < P_{(k-1)j}.$$

		Items				
		1	2	3	...	k
Raw Scores	1	P_{11}	P_{12}	P_{13}	...	P_{1k}
	2	P_{21}	P_{22}	P_{23}	...	P_{2k}
	3	P_{31}	P_{32}	P_{33}	...	P_{3k}
	.	.	.	.	.	.
	.	.	.	.	.	.
	$k-1$	$P_{(k-1)1}$	$P_{(k-1)2}$	$P_{(k-1)3}$	...	$P_{(k-1)k}$

Figure 2. Raw score by item matrix illustrating unidimensionality.

Factor analysis is the method commonly suggested for constructing unidimensional tests (Lumsden, 1961) or to determine whether existing tests are unidimensional. Hambleton et al, (1978) suggest that the factor analytic

approach to test construction outlined by Lumsden be applied in constructing tests which are to be analyzed by latent trait methods.

Hambleton and Traub (1973) performed a principal components analysis on the matrices of tetrachoric item intercorrelations to determine whether three tests were unidimensional before applying the Rasch model in test analysis. In all three cases, the first factor accounted for 20.5% to 31.7% of the total test variance. The researchers proceeded with the analyses as if the tests were unifactorial pointing out, however, that from two to three factors would probably be required to arrive at an acceptable solution for the tests. The best goodness-of-fit of items results obtained, applying the Rasch model and Birnbaum's two-parameter model, were for the most homogeneous test. In a later paper, Hambleton and Cook (1977) stated that the unidimensionality assumption does not hold for most tests, but that a factor analysis could be used to cluster items in a heterogeneous test. Latent trait model item analysis could then be applied to each homogeneous cluster of items.

The factor analytic approach to determining the unidimensionality of a test has been criticized as unnecessary or misleading by other researchers. Factor analysis is itself a series of theoretical models with differing and sometimes conflicting conceptualizations of dimensionality (Rentz & Rentz, 1978). The dimensionality of a set of test items is too dependent upon the types of inter-item correlation coefficients used in the factor analysis. For example, analysis of phi coefficients generally yields more factors than analysis of tetrachoric correlation coefficients (Hambleton & Traub, 1973).

Wright (1977a) cautions against the use of the factor analytic approach, citing the extreme instability of factors over different samples of examinees as a negative feature of the method. It has been pointed out that the best test of unidimensionality may be the direct application of the Rasch model to a group of items (Rentz & Rentz, 1978). In many practical applications of

the Rasch model, items have been successfully analyzed despite the fact that tests were a seemingly heterogeneous, multidimensional set of items. Thus, the Rasch model appears to display robustness with respect to the violation of the unidimensionality assumption inherent in latent trait models.

The assumptions of equal item discrimination and lack of guessing. It has been stated that the Rasch model may not be robust for some data in that the only item characteristic it takes into account is item difficulty, while neglecting to incorporate other possible item characteristics such as item discrimination and guessing (Wright & Panchapakesan, 1969). Thus, an item may actually measure the same trait as the other items with which it is grouped, but it may be rejected in Rasch analysis of the test data because of the model's implied restrictive assumptions that all item discrimination characteristics are equal and that there is minimal guessing involved in answering multiple-choice items.

Two-parameter and three-parameter models incorporating item characteristics for discrimination and guessing exist in the field of latent trait theory. One line of research has been to compare the Rasch logistic model to these other models in situations where the assumptions of equal item discrimination and no guessing have been known to be violated.

The forms of item characteristic curves vary depending upon the number of parameters used in defining them (Hambleton & Cook, 1977). For the Rasch logistic model, the ICC's are non-intersecting, differing only by a translation along the ability scale. These curves reflect the notion that items are held to vary only with respect to their difficulty. Item characteristic curves for the two-parameter model vary in both slope and translation along the ability scale. The variation in slope is a reflection of the fact that some items, with a slope greater than unity, are more discriminating than

other items which have a slope less than one. Item characteristic curves for the three-parameter model differ in slope, translation along the ability scale, and in lower asymptote. The lower asymptote is greater than zero and approaches a hypothetical lower limit which is a factor of the number of alternatives in a multiple-choice question.

In a computer simulation study (Hambleton & Traub, 1971), the one-parameter Rasch model was compared with respect to its effectiveness in determining person ability estimates to Birnbaum's two-parameter and three-parameter latent trait models. Four ranges of item discriminating level (0.0, .2, .4, .8) and three mean levels of guessing (0.0, .1, .2) were generated by means of a computer program. The interaction of varying item discrimination and guessing levels was compared for each model as regarded the model's efficiency in determining person ability levels ranging from -3 to 3. It was concluded that the Rasch model yields comparably efficient estimates of ability when there was no guessing and the range of item discriminations was small. It was found that the Rasch model yielded the least efficient estimates of person ability when low ability 'examinees' responded to a test in which guessing was a factor.

In a subsequent study, Hambleton and Traub (1973) compared the Rasch model and the two-parameter model in an empirical analysis of three different tests of ability. They concluded that both models estimated examinee ability in an equally efficient fashion provided that the number of items in the test was large. The Rasch model was least efficient in estimating person ability levels when the test was short and there was substantial variance in the item discrimination parameters. The numbers of items in the tests analyzed were 20, 45 and 80.

In another simulated study (Dinero & Haertel, 1977), data from the responses of 75 'subjects' in tests of 30 artificially generated 'items' were

fitted to the Rasch model. Five different variances for item discrimination values were generated (.05, .10, .15, .20, .25). For each variance level, three different types of item distribution (normal, uniform, positively skewed) were also simulated. It was concluded that the uniform distribution yielded the poorest fit of items to the model over all levels of item discrimination variance and that there generally was a slight increase in lack of fit of items as the discrimination level variances increased. The researchers concluded, however, that the lack of an item discrimination characteristic in the Rasch model does not generally result in poor fit of items with varying discriminations. They pointed out that the fit of an item to the model probably affects estimates of discrimination, with poor fitting items having larger errors of estimates of their discriminations.

O'Shea (1979) cited a simulated study by Cartledge (1976) in which the researcher found that items with discrimination values ranging from .9 to 1.1 were found to fit the Rasch model, and that the model continued to display robustness when the item discriminations varied from .8 to 1.2.

In a simulated study, Panchapakesan (1969) found that the items with extremely deviant discriminations did not fit the model, but that items with discrimination values ranging from .8 to 1.25 displayed fit to the model. Using a sample size of 500 for a 20-item simulated test containing items with discrimination values ranging from .8 to 1.2, Panchapakesan found that the mean square for overall fit of items was not large enough to be significant. She concluded that the Rasch model displayed robustness when the assumption of items with equal discriminations was not met.

There is evidence from simulated studies employing exactly equal item discrimination values, that the standard deviation for estimates of discrimination values are as large as .2 (Wright & Mead, 1978). The inference which might possibly be drawn from this and the previous studies cited, is that a

range in discrimination values for items from .6 to 1.4 might be acceptable (O'Shea, 1979).

Since guessing is a problem inherent in the use of multiple-choice test items, it might be well to set some lower limit to scores which are allowed to enter into data analysis using the Rasch model. This would compensate for the absence of a guessing characteristic in the model. One approach suggested by Wright and Mead (1978) is to set the lower limit somewhat above the random guessing level. This approach is often the one followed to control for guessing in conventional item analysis. Panchapakesan (1969) in simulated studies involving a 20-item test composed of five-alternative multiple-choice items, concluded that the effect of guessing was negligible provided that the 'examinees' were of high ability. For situations in which this is not the case, she proposed the following formula for eliminating examinees from the calibration sample:

$$\underline{r} = \frac{\underline{k}}{\underline{m}} + 2\sqrt{\frac{\underline{k}(\underline{m} - 1)}{\underline{m}^2}},$$

where $\underline{k}$ = the number of items in the test,

$\underline{m}$ = the number of alternatives per item, and,

$\underline{r}$ = the score below which examinees are eliminated from the calibration sample.

In the formula above, $\underline{k}/\underline{m}$ is the expected score based on random guessing and $(\underline{k}(\underline{m} - 1)/\underline{m}^2)^{0.5}$ is the standard deviation of that score.

Estimation of Item Parameters for the Rasch Logistic Model

Once the form of the item characteristic curve for the Rasch model has been specified and the necessary assumptions of the model have been met, it is possible to arrive at estimates of the parameters $\underline{a}_s$ and $\underline{d}_i$. Estimators for the Rasch model have been developed which have been shown to be sufficient, consistent, efficient, and unbiased (Wright, 1977a). Several computer

programs exist for the computation of the parameters, as well as a method by Cohen (1976) which can be used to arrive at approximate estimates by hand.

The data necessary for the estimation of the K item parameters and the N person parameters for a test consists of the matrix of K items on the test (1, 2, ..., K) arranged in the columns and all possible scores other than zero and K (1, 2, ..., K - 1) arranged in the rows. Perfect and zero scores are removed from the data base because such scores do not yield finite estimates of ability for persons obtaining them (Wright & Douglas, 1977a). Similarly, items which all examinees get correct provide no differential information about actual differences in difficulty of these items.

Rasch (1966) showed that the column and row totals for items and persons in a test score matrix were jointly sufficient statistics for the simultaneous estimation of the person ability parameter, a_s, and the item difficulty parameter, d_i. He also deduced from the model that the column and row totals of the matrix could be used as separate and independent estimates of the parameters d_i and a_s respectively.

Two computer programs for obtaining parameter estimates, LOG, which uses an unweighted least squares approach, and MAX, which uses a maximum likelihood procedure, were outlined by Wright and Panchapakesan (1969). They favoured the maximum likelihood procedure as it yielded better estimates of a_s and d_i as well as better approximations of the standard errors of the estimates. They pointed out that when N was large, and the ability range of the examinees was wider than the difficulty range of the item parameters, the item estimates obtained by LOG and MAX were equivalent.

Andersen (1973) criticized the Wright and Panchapakesan approach because it yielded unconditional estimates of a_s and d_i. He showed that, as N increases, the number of ability parameters increases without bound leading to biased estimates of the item parameters.

Wright and Douglas (1977b) outlined two computer programs, FCON, the full conditional procedure based on Andersen's (1973) solution of the conditional likelihood equations, and ICON, an incomplete conditional procedure for estimating the a_s and d_i parameters. They found the former procedure impractical for use with more than 10 or 15 items because it did not converge rapidly with numbers of items greater than this. They found that the ICON procedure converged more rapidly, but that biased estimates of the item parameters resulted when there were more than 20 or 30 items in a test. Since ICON required a correction factor for bias, the researchers concluded that the most fruitful approach would be to utilize the original unconditional procedure, UCON, (Wright & Panchapakesan, 1969) with a correction for bias since this was the most practical program for implementation. They determined that a correction factor of $(K - 1)/K$, when applied to the biased estimates of the item parameters, yielded values very close to the correct ones (Wright & Douglas, 1977a). This correction factor for item parameter bias is incorporated into the BICAL computer package prepared by Wright and Mead (1978).

Wright and Douglas (1977a) also presented an approximate procedure, PROX, which requires the assumptions that person abilities and item difficulties are normally distributed. They found that the calculation time using PROX for large numbers of items was approximately 40 times faster than the time it took for UCON to converge. Hambleton et al, (1978) cautioned that PROX requires further investigation as it is an approximate procedure and the extent to which departures from the assumption that item difficulties and person abilities are normally distributed affect the estimates is not well understood. PROX is an option available in the BICAL computer package.

Gustafsson (1980) described a procedure for conditional estimation of item parameters for the Rasch model which provides rapid computation of nu-

merically stable estimates even for long tests consisting of 80 to 100 items. He concluded that for tests of this length a conditional approach to item parameter estimation is superior to an unconditional approach in that powerful tests of goodness-of-fit of data to the Rasch model, with known statistical properties, can be devised.

Standard error of estimate for item difficulties. The Rasch method of test analysis provides a measure of the standard error of estimate associated with the item difficulty parameters. The standard error of estimate for item difficulty in the Rasch model is minimized when there is a high correspondence between item difficulties and the mean ability of persons in the sample. The best estimates of item difficulty are obtained for items on which the probability of passing the item or failing the item is equal or nearly equal for the maximum number of persons in the sample (Whitely & Dawis, 1974).

The converse is true in the usual conceptualization of item variance employed in classical test theory. Item variance, $\underline{s}_{\underline{i}}^2 = \underline{p}_{\underline{i}}\underline{q}_{\underline{i}}$, is the highest when there is a 50% chance of passing or failing an item, i.e., when $\underline{p}_{\underline{i}} = \underline{q}_{\underline{i}} = .5$, and is minimized as items become easier or more difficult. This situation seems paradoxical, especially in the case of very difficult multiple-choice items where guessing might be most prevalent.

While some investigators maintain that the Rasch model yields item-difficulty estimates which are invariant over samples differing in ability, this is not the case for standard errors of estimate for item difficulties (Whitely & Dawis, 1974). The standard error of item difficulty is sample-specific, being strongly influenced by the relationship between item difficulty and the distribution of ability in the sample (Wright, 1977a).

Standard error of estimate for person abilities. Use of the Rasch model also provides separate errors of measurement for each of the person ability parameters. The most precise measurements of person ability are obtainable

from central scores on a test and error in estimating person ability is maximized as scores become more extreme in either direction. Thus the most precise estimates of person ability will be obtained if tests are composed of items closely complementing the abilities of the examinees (Whitely & Dawis, 1974).

The standard errors of estimates for person abilities obtained using the Rasch model behave in the opposite manner from the conventional standard errors of measurement for randomly parallel tests consisting of items randomly sampled from a given domain. In the conventional case, the standard errors of measurement for scores become minimal as the scores depart in value from the central cases. They illogically become zero at scores of zero or 100%, where the estimation of person ability is actually extremely imprecise (Wright, 1977a).

Again, claims have been made that the Rasch model provides invariant person ability estimates which do not depend upon the particular set of items chosen to measure ability (Wright, 1977a). However, the precision of the person ability estimates will vary directly with the extent to which the difficulty of the items chosen to measure that ability complements the actual ability of the persons being measured.

Testing Item Fit to the Rasch Model

Several approaches have been used in evaluating the fit of test items to the Rasch model. Rasch (1960) originally employed a simple graphical approach to approximately determine how well his model applied to test data (O'Shea, 1979). Rasch grouped examinees into five ability groups ranging from high to low on the basis of raw scores obtained in a test. His approach to testing the fit of the data to his model was to plot two-way graphs of item difficulty for all pairs of adjacent score groups. Rasch then made a visual confirma-

tion of the fit of his data by noting whether the four lines of best fit were approximately parallel and had a slope approaching unity. If score groups are of adequate size, slopes of items can be adequately estimated with unit slope being the criterion for fit. Procedures using all score groups probably do not yield good estimates for extreme score groups. The BICAL procedure collapses score groups into larger entities. Consequently persons having extreme scores are assimilated into larger groups which probably provide more stable estimates (Rentz & Rentz, 1978).

A more precise procedure was to use a statistical approach to test the significance of the departure of regressions from a unit slope (Anderson, Kearney & Everett, 1968). A group of persons taking a test was divided into six ability groups. For each item, a difficulty estimate specific to each score group was computed. A mean value for all item difficulties was obtained for each ability group as well. The six difficulty estimates for each item were then regressed on the six means of all difficulty estimates for each score group. A t-value was obtained for each item to determine whether the regression of the six item difficulty estimates on the six overall difficulty means differed statistically from a unit slope. Two probability levels were used for the t-tests. The test using a probability of .05 was the more stringent test of fit to the model as more items were rejected using this criterion than were rejected when a probability value of .01 was used.

Test of fit of items to the Rasch model by determining a chi-square statistic, or some variation thereof, has become the common method (Rentz & Rentz, 1978). The general procedure was outlined by Wright and Panchapakesan (1969) and involves the comparison of data obtained with what is expected given the model. The starting point is a raw score matrix for a test with k items and m score groups with no empty cells, or a matrix in which score groups have been collapsed into a smaller number of groups, j. For each cell,

the standard deviate is the difference between the observed value and the estimated expected value of the probability of success on the item. Given these values, a suitable statistic for testing the overall fit of the entire group of items to the model is the chi-square statistic--the sum of all squared standard deviates with degrees of freedom $(k - 1)(m - 1)$, or in the case of grouped scores $(k - 1)(j - 1)$. The fit of each individual item to the model can also be evaluated by summing the squared standard deviates over the score groups with $(m - 1)$ or $(j - 1)$ degrees of freedom. This chi-square statistic was used by Fryman (1976) who eliminated items which had a probability of fit to the model by chance of less than .05.

The quotient of the chi-square divided by the appropriate degrees of freedom yields the mean square fit statistic. Care must be taken in the interpretation of the MS fit statistic since, as with all statistical tests of significance, sufficient power is invariably gained to reject the hypothesis--in this case a question of whether an item fits a particular model--if a large enough sample is used (Rentz & Rentz, 1978). The MS fit statistic will lead to the elimination of items in all instances if very large samples are used. Consequently, MS fit statistics should be considered as a group in relation to each other. Outliers which do not fit the model should be spotted instead of setting absolute values of the MS fit statistic for selection/rejection of items. If the sample is a very large one, items with MS fit values as high as 10 may still fit the model if this value does not differ greatly from the MS fit values shown by the items in the whole group (Rentz & Rentz, 1978).

Hashway (1977) used the MS fit statistic together with the item discrimination indices provided by the Rasch model to reject items which had a MS fit greater than 2.2 and a Rasch discrimination index outside the range .8

to 1.2. O'Shea (1979) established similar criteria for rejecting nonfitting items, eliminating items for which the MS fit exceeded unity by four or more standard errors of the MS fit, and for which the Rasch discrimination index was less than .7. The Rasch discrimination index provided by computer programs such as BICAL is related to the conventional point-biserial coefficient (O'Shea, 1979). Indeed, hybrid approaches to item analysis, using the Rasch MS fit statistic together with the conventional point-biserial correlation coefficient have been employed. Forbes (1976) eliminated items having a point-biserial item-total score correlation less than .25 and a mean square fit exceeding 2.5. These latter studies indicate that researchers have used fit criteria other than chi-square related statistics in determining the fit of test items to the Rasch model.

The question of sample size in application of the Rasch model. There is some question as regards the size of the sample required to test the fit of items to the Rasch model. In an early study using the Rasch procedure, samples of 446, 129, and 279 subjects were considered too small for application of the technique (Anderson et al, 1968). In another study, the investigators pointed out that correlations between item difficulty estimates obtained from 89-subject samples may have been low because of the small sample size (Tinsley & Dawis, 1975). They pointed out, however, that estimates from a larger sample of 269 subjects also did not correlate highly.

Whitely (1977) has taken up the issue of sample sizes used in testing fit of items to the Rasch model. She points out that the test of fit is a statistical test and like all statistical procedures requires a sample of a reasonable size to discriminate between the small differences that may exist between good and bad items. Whitely stated that an N as large as 800 may fail to detect sizable differences between the quality of fit of test items.

In reply to Whitely, Wright (1977b) argued that the sample size required to test fit of items to the Rasch model is simply dictated by the precision of the calibration of items which the test builder wishes to achieve. Wright felt that a sample size as small as 100 may provide useful information in certain test building situations.

In an empirical investigation of the matter of sample size, it was found that an N between 200 and 300 was sufficient for field testing items to be subjected to analysis using the Rasch model, with a sample size of 200 providing information nearly as accurate as a sample size of 300 (Forster et al, undated).

Test Analysis and Development Using the Rasch Model

Once item parameters have been estimated and fit of the items to the Rasch model has been evaluated, tests may be constructed and calibrated to obtain an ability-scale value corresponding to each raw score on the test.

The Rasch model has been applied in test construction and analysis in the school content areas of reading (Forster et al, undated; Rentz & Bashaw, 1977) and mathematics (Forster et al, undated; Hashway, 1977; O'Shea, 1979) at the elementary level. School achievement tests investigated have typically been multiple-choice, but it is probable that the model will accommodate items with other modes of response provided the tasks can be dichotomized into pass or fail scores (Rentz & Rentz, 1978).

There is evidence that the Rasch procedure is somewhat harsher in identifying unsuitable items than is the conventional procedure of item analysis. Between 5% to 10% of the items in some tests have been rejected by Rasch criteria after conventional analysis has been applied (Rentz & Rentz, 1978). However, this would naturally depend upon the values of the rejection criteria chosen. When the range of item discriminations is wide, Rasch procedures and conventional procedures will tend to differ in the rejection of items.

Rasch procedures eliminate items at both extremes of the discrimination value spectrum while conventional procedures eliminate only low discriminating items. However, studies comparing the characteristics of items selected for tests using conventional and Rasch procedures have failed to pinpoint substantial differences between items selected/rejected by the two methods (Hambleton et al, 1978).

Conventional item analysis and the Rasch procedure were used to develop two homogeneous subscales from the 49-item MacAndrew alcoholism scale administered to 373 university students (Douglass, Khavari & Farber, 1979). Conventional item analysis yielded a subscale of 29 items from the original scale each having item-total test correlations of .2 or greater. The subscale constructed using the Rasch method consisted of 36 items remaining after items having a chi-square value failing to fit the model with a probability less than or equal to .001 were discarded.

The investigators found substantial differences in the total number of items chosen by each method, specific items chosen by each method, and the relative fit of items depending upon which procedure was used. They were also able to arrive at different definitions of the traits measured by the conventional subscale and the Rasch subscale.

The Rasch raw score scales and the Rasch ability estimate scales correlated more highly with the original 49-item test than did the conventional raw score scales. However, evaluation of the correlations with a t-test for dependent correlations yielded no significant differences. The conventional subscale, the Rasch subscale, and the original 49-item full scale all had equally low correlations with an external criterion, the Khavari Alcohol test for assessing alcohol consumption.

The researchers concluded that neither the conventional subscale nor the Rasch subscale had been effective as a purified subscale derived from an al-

ready well-constructed homogeneous scale. They felt that the conventional method should remain the method of choice for test construction since it was more easily understood and was less costly as regards computer time.

Ryan and Hamm (1976) used the Rasch procedure in an attempt to increase internal consistency, as measured by KR 20, and to clarify the factor structure of the items on eight teacher-made tests in educational research. The number of items in the tests varied from 15 to 50. All items were four-option multiple choice questions. Two subscales were constructed from each of the original tests, a subscale of items remaining after deletion according to Rasch criteria, and a subscale consisting of items loading on the first factor in a principal components analysis of the original test. A KR 20 reliability coefficient was calculated for each of the eight original tests as well as for the two tests derived from each original test in order to determine the effect which item selection approach had on test internal consistency. Since the two subscales developed from each test consisted of fewer items than the original test, the researchers applied the Spearman-Brown formula to the KR 20 value for each subscale so that the values would be comparable for a test as long as the original test.

With the Spearman-Brown formula applied to the KR 20 values it was found that all subscales constructed using the Rasch approach, and seven of the subscales constructed using factor analysis, yielded a KR 20 coefficient superior to that of each related original test. Omitting the Spearman-Brown correction formula, seven of the Rasch subscales and two of the factor analyses subscales still had KR 20 reliabilities superior to that of the original tests. It should be noted that the Rasch approach retained more items from the original tests than did the factor analytic approach.

The effects on the factor structure of the tests by deleting items were compared by performing a factor analysis on the original scale, the Rasch

subscale, and the factor analysis subscale of each test. The proportion of variance accounted for by the first factor was highest for tests constructed using first factor items. The first factor of the Rasch subscales accounted for only a slightly higher proportion of the variance than did the first factor on the original tests.

CHAPTER III

DESIGN OF THE STUDY

The purpose of the study was to objectively apply conventional and Rasch item analysis procedures to a pool of test items and to compare the specific nature of the items selected/rejected as well as the overall characteristics of tests resulting from the application of the two approaches. The study concerned itself with the practical application of two different statistical item analysis procedures. It was not concerned per se with ensuring that the resulting tests possessed content validity which would have involved subjective judgements regarding the items to be included.

Item Pool

The item pool used in this study was a modified version of the 75-item pool used for the development of the Grade 4 Science Assessment pilot tests. First, the item pool used in this study contained an additional item which had been added to the final version of the test used in the assessment. Second, three items which all referred to the same graphical material and appeared in consecutive order were separated. Although these items were not chained, it was felt that keeping the items clustered together might violate the assumption of local independence for the Rasch model. Third, item 22 from pilot test booklet A and item 30 from pilot test booklet B were rewritten. The former item did not have a stem compatible with its options and the latter item had two correct options. Fourth, the final version of the 68 items used in the actual assessment was adopted. And fifth, instead of using two separate booklets, all 76 items were included in one test. The goal areas and the items assigned to them in the 76-item version of the item pool

used in the present study were the same as those used for the original item pool and are shown in Table 2. A copy of the test booklet used is provided in Appendix A.

The Sample

Sample Size

The initial minimum sample size was set at $n = 500$. An n of 500 would satisfy the apparent minimum requirement of at least 200 to 300 persons needed for Rasch item analysis as discussed in Chapter II. A sample of 500 would

Table 2

Assignment of Items to Goals

Goal 1--Understanding Concepts (30 items)
1, 3, 4, 5, 6, 7, 10, 12, 13, 19, 20, 21, 24, 27, 28, 30, 34
38, 40, 42, 45, 46, 51, 52, 53, 54, 55, 58, 67, 72
Goal 2--Skill in the Processes of Science (32 items)
Classifying: 9, 11, 44, 49, 65, 76
Observing: 16, 17, 23, 35, 60
Communicating: 14, 31, 32, 59, 70, 71
Measuring: 8, 18, 26, 29, 37, 41, 48, 63, 73
Combined Skills: 22, 33, 36, 56, 62, 74
Goal 3--Application (5 items)
39, 43, 50, 66, 69
Goal 4--Safety (5 items)
2, 25, 61, 68, 75
Goal 5--Scientific Literacy (4 items)
15, 47, 57, 64

also satisfy the requirement of 300 to 500 subjects suggested by Henryssen (1971) for a meaningful conventional item analysis. Furthermore, factor

analysis was employed in the present study as a method of determining test homogeneity. Where factor analysis of tests is involved, it is suggested that there be a sample consisting of at least 10 subjects per item (Nunnally, 1967). The two largest subtests used in this study consisted of 30 items (Concepts) and 32 items (Processes). For factor analysis of a 32-item subtest a sample size of 500 subjects exceeds this guideline by 180 subjects.

Obtaining the Sample

In order to exercise greater control over the distribution and collection of test booklets, and since available time and funds were limited, the sample selected was restricted to the Okanagan Valley of British Columbia. Accordingly, permission was sought from the superintendents of the six school districts in this area to administer the tests to the grade four students in their districts.

Shortly after mid-October, 1979, the superintendents were sent a letter (see Appendix B) in which the purpose of the study as well as the nature of the test to be used was described. It was stressed that the purpose of the study was to compare methods of item analysis and not to compare individual student, school, or district achievement. The strictest confidentiality concerning such data was assured. A stamped, self-addressed return envelope was included with each letter.

Five superintendents granted permission to contact the elementary school principals in their districts[1]. Accordingly, in November 1979, 59 elementary school principals and head teachers were contacted by letter (see Appendix B). The letters contained information similar to that covered in the

---

1. The sixth superintendent answered affirmatively, but his letter was not received until after the letters to principals had been sent and just shortly before the testing period was to commence. Consequently, students from this district were not tested.

letters to the superintendents. A testing period during the week of December 3 - 7, 1979, was suggested. It was emphasized that the tests would be computer scored. The letter included a tearoff sheet on which principals could indicate their decision to allow grade four classes in their schools to take part in the study or not. The tearoff sheet also included a space for the number of copies of the test required for each school if the decision to take part in the study was positive. Stamped, self-addressed return envelopes were again provided.

Of the 59 principals approached, 31 replied. Nineteen principals granted permission to conduct the testing in their schools. Of the 12 principals who replied negatively, several provided explanations for their decisions. In one case, the school was a primary school with K - 3 classes only. Another small school did not wish to take part in the study because they had only 11 grade four students distributed in split grade three/four and grade four/five classes. Two principals mentioned that their schools would be heavily involved in seasonal projects at the time of testing. In one school the principal was on medical leave of absence and the person in charge wished as few complications as possible during the regular administrator's absence. Also, schools in the target group were involved in a science pilot test project at the grade four to seven level being conducted by the Educational Research Institute of British Columbia (Educational Research Institute of British Columbia, 1980).

Students were tested in 15 of the 19 schools in which permission to test had been granted. These schools enrolled approximately 550 grade four students and were all located in a compact geographical area in the southern part of the Okanagan Valley, thereby facilitating the data collection. Three schools had a grade four enrolment of 20 or less; seven had a grade four enrolment of between 21 and 40 pupils, and five had a grade four enrolment of between 41 and 60 pupils.

Testing

Test packages were delivered to the 15 sample schools during the week prior to December 3 - 7, 1979. A covering letter to the principal of each school gave instructions as to when the tests were to be administered and stated that completed tests would be collected the week after testing (see Appendix A).

The test packages included sets of directions for the teachers who were to administer the tests (see Appendix A). The directions included an example question which was to be put on the blackboard for all students to see. The teachers were asked to ensure that all students had a centimeter ruler which was required for the completion of question 37. It was stressed that students retain their test booklets if testing were interrupted by a rest period. An identification procedure to ensure this was suggested. The procedure preserved the child's anonymity. Teachers were free to devise their own identification procedures if they wished. Teachers were asked to encourage students to attempt all questions and to use the 'I don't know' option only if they did not know the answer. To avoid problems associated with speeded tests in conventional and Rasch item analysis, teachers were asked to allow sufficient time for at least 95% of the students in their classes to complete all items. The teachers administering the tests were asked to provide pertinent written comments about the test if they wished.

Data Preparation

A total of 527 completed test booklets was collected. Preliminary inspection of the test booklets indicated that students generally had sufficient time to complete the test. To alleviate potential effects due to speededness upon subsequent analyses, booklets which satisfied the following criteria were eliminated:

1. If a test booklet had more than five unanswered items at the end of either of the two test sections the booklet was eliminated on the basis that the pupil apparently did not have enough time to complete one or both sections of the test.
2. If a test booklet had more than 10 items which were unanswered or for which more than one response had been indicated, the booklet was eliminated from the analysis. The basis for this decision rule was that a student who had more than ten such items in his/her booklet may have been too hurried to give full consideration to each item, or may have had a careless attitude towards completion of the test.

In cases where students had not answered isolated items throughout the test, or had provided more than one answer to an item, the 'I don't know' option was keypunched as the response for such items to a maximum of ten items per booklet. It was felt that students who did not answer a particular item or who were confused regarding the number of possible answers to an item likely did not know the answer to that item.

Of the 527 test booklets obtained, 11 were eliminated on the basis of the criteria above. Data for the remaining 516 subjects were keypunched and verified by comparing the keypunched data to the actual responses in the student booklets. The overall error rate was .16 percent. All errors were corrected.

Data Analysis

Restriction of Analyses to Concepts and Processes Subtests

Preliminary analyses indicated that treatment of small subtests containing four or five items became problematical when using the BICAL computer program. Normally BICAL employs six score groups in computing item fit to the model. If a subtest consists of five items, this leaves only four inte-

gral values--one, two, three, and four--which can be used to strike score groups for the analysis. Five, representing a perfect score, is not used in the analysis for reasons discussed in Chapter II. While the number of score groups used in BICAL can be varied, it would seem that Rasch indices obtained for a cluster of five items might be quite unreliable because of the restricted number of score groups involved in their calculation, and the vastly disproportionate numbers of subjects which might occur in each group.

For the above reason, it was decided to focus attention on the two larger subtests involving science concepts and the processes of science. As shown in Table 2, these subtests consisted of 30 items and 32 items, respectively. It was felt that the most reliable and useful comparisons could be made by analyzing items from these larger subtests. Consequently, the items in the Application, Safety, and Scientific Literacy subtests were excluded from the subsequent analyses.

Tests for Unidimensionality

As mentioned in Chapter II, it is generally felt that latent trait test models such as the Rasch model are applicable only to tests which are unidimensional (Hambleton & Traub, 1973). Accordingly, the following steps were taken to determine whether the Concepts and Processes subtests were unidimensional:

1. Cronbach's coefficient alpha. Cronbach's coefficient alpha was .76 and .83 for the Concepts and Processes subtests, respectively. When the Processes subtest was stratified into five groups of items, each group relating to one of the objectives--Classifying; Observing, Communicating, Measuring, Combined Processes--stratified coefficient alpha was .79. These relatively high values suggested that each of the two subtests was essentially unidimensional.

2. Factor analysis. Factor analysis was performed on the Concepts and Processes subtests. A short computer program was written to render the original test responses into binary scored form--0 for an incorrect answer and 1 for a correct answer. Tetrachoric correlation coefficients were then computed among the items for each subtest. The resulting correlation matrices were subjected to a principal components analysis using the AGFAP factor analysis program (Hakstian & Bay, 1973). The number of factors retained for transformation was determined by two criteria--a) the number of eigenvalues greater than unity, that is, the Kaiser-Guttman criterion, and b) the scree test as described by Gorsuch (1974).

The results from the principal components analyses on each of the two subtests are presented in Table 3. As shown, the number of factors to be extracted was problematical. Applicaton of the Kaiser-Guttman criterion suggested a greater number of factors than the scree test. To further clarify the number of factors, the solutions were transformed using the Harris-Kaiser method of oblique transformation with $c = 0$ (independent cluster solution). Pattern coefficients greater than $|.40|$ were considered to be salient. In the case of both subtests, for each number of factors retained, groups of items with salient loadings did not lend themselves to meaningful interpretation as being homogeneous clusters of items measuring an obviously definable trait. Because factor analysis did not separate the items into clusters measuring clearly different traits it provided additional evidence for the unidimensional nature of the two subtests used in this study. It was felt that the identification of more than one factor after extraction was attributable to problems associated with dichotomous variables in producing spurious factors (Gorsuch, 1974).

In a third investigation a principal components factor analysis of the matrix of Pearson product-moment correlations among the five item clusters in

the Processes subtest was performed. As shown in Table 3, the results of this analysis provided further evidence for the unifactorial nature of the Processes subtest.

Table 3

Factor Analyses of Subtests

Subtest	Variables	Type of Correlation Coefficient	Factors Extracted Eigenvalues >1 Rule	Scree Test
Concepts	30 items	tetrachoric	10	2, 6
Processes	32 items	tetrachoric	8	4
Processes	5 item clusters	Pearson product-moment	1	

Item Analysis and Test Construction

The 30 Concepts items and the 32 Processes items were treated as two subtests. Items were analyzed separately in each subtest. Using both the conventional and Rasch procedures, Concepts and Processes subtests were built according to the following seven sets of criteria:

1. Conventional Approach

   The items in each of the two subtests were subjected to a conventional item analysis using the LERTAP (Nelson, 1974) computer program. Items were discarded if they failed to meet one or more of the following criteria;

   a. The difficulty index, $\underline{p}$, must lie within the range $.20 \leq \underline{p} \leq .80$.

   b. The corrected item-subtest point-biserial correlation coefficient must be equal to or greater than .20.

   c. All distractors must have an uncorrected item-subtest

point-biserial correlation coefficient less than 0.

d. All distractors must have a nonzero difficulty index.

2. Rasch Approach I (Panchapakesan)

Examinees with scores below the guessing level set using Panchapakesan's formula were excluded from the calibration sample for that subtest. Items were discarded if they failed to meet one or more of the following criteria:

a. The mean square fit must be less than one standard deviation above unity.

b. The discrimination index must be within one standard deviation of the mean item discrimination.

c. The difficulty index must be within one standard deviation of the mean item difficulty.

3. Rasch Approach II (Panchapakesan)

Panchapakesan's formula was again applied in determining the composition of the calibration sample. For this analysis, criterion c as applied in Rasch Approach I above was dropped.

4. Rasch Approach III (Panchapakesan)

Application of Panchapakesan's formula again determined the make-up of the calibration sample. Criteria b and c as applied in Rasch Approach I above were dropped.

5. Rasch Approach IV (Random)

This analysis used the same criteria as Rasch Approach I above. In determining the calibration sample, examinees with scores below the random guessing level for each subtest, set using the $\underline{k}/\underline{m}$ formula, were excluded.

6. Rasch Approach V (Random)

This analysis used the same criteria as Rasch Approach II above.

The composition of the calibration sample was determined by applying the random guessing level formula.

7. Rasch Approach VI (Random)

This analysis used the same criterion as Rasch Approach III above. Application of the random guessing level formula determined the make-up of the calibration sample.

Correction of item-subtest point-biserial correlations. For shorter tests it is advisable to correct item-test correlation coefficients for spuriousness. An appropriate correction formula for use with the point-biserial correlation coefficient is Guilford's (1953) formula as discussed by Berk (1978):

$$\underline{r}_{\underline{i}\underline{r}} = \frac{\underline{r}_{\underline{i}\underline{t}}\underline{s}_{\underline{t}} - \underline{s}_{\underline{i}}}{\sqrt{\underline{s}_{\underline{i}}^2 + \underline{s}_{\underline{t}}^2 - 2\underline{r}_{\underline{i}\underline{t}}\underline{s}_{\underline{i}}\underline{s}_{\underline{t}}}},$$

where $\underline{r}_{\underline{i}\underline{r}}$ = the correlation of an item with the remainder of the total score after removal of the item,

$\underline{r}_{\underline{i}\underline{t}}$ = the point-biserial correlation obtained between the item and the total score with the item included,

$\underline{s}_{\underline{i}}$ = the standard deviation of the item, $\underline{p}_{\underline{i}}\underline{q}_{\underline{i}}$, provided the items are dichotomously scored with weights of one for correct items and zero for incorrect items, and

$\underline{s}_{\underline{t}}$ = the standard deviation of the total test score.

When the item-test correlations are high--.40 or greater--application of the correction formula does not appear to be necessary. However, the status of items having point-biserial correlations close to the cutoff point--$\underline{r}_{\underline{pb}}$ = .20 in this study--is often changed when the correction formula is applied. For example, consider an item having an $\underline{r}_{\underline{pb}}$ of .25 with the total test and a p-value of .31. If the total test has a standard deviation of 4.86, the

corrected $\underline{r}_{pb}$ is .16. Application of the correction formula would indicate rejection of the item using the criteria of this study.

Elimination of subjects from the Rasch analyses. Two methods were used to eliminate subjects from the calibration samples used in the six Rasch approaches described above. For the first three (Approaches I, II, and III) the minimum score for inclusion of subjects in the calibration sample was determined by using Panchapakesan's (1969) formula:

$$\underline{r} = \frac{\underline{k}}{\underline{m}} + 2\sqrt{\frac{\underline{k}(\underline{m} - 1)}{\underline{m}^2}},$$

where $\underline{k}$ = the number of items,

$\underline{m}$ = the number of alternatives per item, and

$\underline{r}$ = the score below which examinees are eliminated.

Panchapakesan's formula is based on the random guessing score, $\underline{k}/\underline{m}$, and takes into account the variance attributable to random guessing, $(\underline{k}(\underline{m} - 1))/\underline{m}^2$. Since the number of options for items in the tests varied from four to six, the minimum score for inclusion of examinees in the calibration sample was determined as a composite. The minimum scores obtainable by guessing on the four-, five-, and six-option items were computed separately and then added to obtain the total minimum score which was rounded up to the nearest whole number.

For the last three Rasch approaches outlined above (Approaches IV, V, and VI), the minimum score for inclusion of subjects in the calibration sample was determined by application of the less conservative formula used to find the score obtainable by random guessing on multiple-choice questions. This approach has been suggested by Wright and Mead (1978). The minimum score is obtained by calculating the expected score based on random guessing, $\underline{k}/\underline{m}$. Again, the score for each subtest was determined as a composite and the minimum score was then rounded up to the nearest whole number. Table 4

provides a summary of the number of subjects whose tests were used in each of the item analyses approaches. As can be seen, there is a large reduction in the number of subjects entering the Rasch analyses in which Panchapakesan's formula was applied. The numbers, however, are above the range of 200 to 300 subjects suggested for meaningful application of the Rasch method in item analysis.

Table 4

Number of Subjects in Each Item Analysis

Analysis	Concepts Subtest	
	Minimum Score	Number of Subjects
Conventional	--	516
Rasch I, II, III	13	394
Rasch IV, V, VI	7	501
	Processes Subtest	
Conventional	--	516
Rasch I, II, III	13	462
Rasch IV, V, VI	7	508

Subtest Comparisons

The above item analysis approaches yielded a conventional Concepts subtest, a conventional Processes subtest, six Rasch Concepts subtests, and six Rasch Processes subtests. The subtests constructed were subjected to a series of comparative analyses as follows:

1. Percentage of items rejected by each approach. The percentage of items discarded from the Concepts and Processes item pools using the criteria

in each of the approaches was computed.

2. Percentage overlap of rejected items. For each possible pair of Concepts subtests constructed using different approaches, there was a common group of items rejected by both approaches. This was also true of each possible pair of Processes subtests constructed. To determine the percentage overlap of items commonly rejected by two approaches, the number of commonly rejected items was divided by the number of commonly rejected items plus the number of items uniquely rejected by each of the approaches. The quotient was then multiplied by 100. For instance, the number of items commonly rejected by the Conventional Approach and Rasch Approach I for the Concepts subtest was 10. Two items were uniquely rejected by the Conventional Approach and six items were uniquely rejected by Rasch Approach I. Thus, the percentage overlap of items rejected by the two approaches for the Concepts subtest was computed as follows--(10/(10 + 2 + 6)) x 100 = 56%. The percentage overlap of rejected items was computed for each possible pair of approaches for the Concepts subtests and for the Processes subtests.

3. Rasch elimination of items not meeting conventional criteria. Since the Rasch model is a relative newcomer to item analysis, it was decided to obtain a measure of the ability of Rasch item analysis to reject items which would be considered unsuitable because they did not meet conventional criteria of item selection. Accordingly, the percentage of Rasch rejected items in each of the Rasch Approaches I to VI which also displayed problems related to conventional criteria was computed by dividing the number of these items by the number of items rejected in the Conventional Approach and multiplying this quotient by 100. For instance, application of the p-value criterion to the Concepts item pool in the Conventional Approach led to rejection of four items. Rasch Approach II was successful in identifying only one of these items. Therefore, the percentage success for Rasch Approach II in identify-

ing Concepts items displaying p-value problems was 1/4 x 100 = 25%. These percentages were computed for each type of conventional item problem, for each combination of conventional item problems, and for the total number of items in the Concepts and Processes item pools displaying conventional problems. Percentages were obtained for each Concepts and Processes subtest built using a Rasch approach.

4. Rasch specific items. For each of the Rasch subtests constructed the percentage of items discarded solely on the basis of Rasch criteria was computed. The total number of items uniquely discarded by the Rasch criteria was divided by the total number of items discarded by the approach and the quotient was multiplied by 100. Thus, the denominator in the computation was the sum of the number of items uniquely rejected by Rasch criteria and the number of items rejected by Rasch criteria which also displayed problems related to conventional criteria. For instance, in Rasch Approach IV, seven items were uniquely discarded by Rasch criteria from the Concepts item pool. An additional nine items, which also displayed problems related to conventional criteria were also discarded by Rasch Approach IV. Therefore, the percentage of Rasch-specific items discarded from the Concepts item pool by Rasch Approach IV was $(7/(7 + 9)) \times 100 = 44\%$.

5. Test characteristics. Each Concepts and Processes subtest resulting from the seven different item analysis procedures was 'readministered' to the original 516 subjects used in the seven approaches. Each subtest was subjected to analysis using the LERTAP computer program. Data obtained for each subtest included test means, standard deviations, and range of scores, as well as Hoyt estimates of reliability and standard errors of measurement.

6. Correlation of subjects' scores on subtests. In order to determine whether the seven Concepts subtests were ordering subjects in a similar fashion, correlations were obtained for the 516 subjects' scores on each possible

pair of Concepts subtests. Similar correlations were computed for subjects' scores on each possible pair of Processes subtests. The scores used were the subjects' arcsine transformed proportional scores on each of the subtests. Because the number of items on each subtest varied it was necessary to convert the subtest scores of individuals to proportions in order to obtain meaningful comparisons. Since proportions form a binomial rather than a normal distribution, it was necessary to perform an arcsine transformation on each proportional score before proceeding with the correlations. The correlations were obtained using the PEARSON CORR program in the SPSS computer package (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975).

7. Comparison of subtest means. Correlated t-tests were performed on all possible pairs of Concepts subtest means, and on all possible pairs of Processes subtest means. Again, arcsine transformations of the 516 subjects' proportional scores on each subtest were used in computing the means of each subtest. The T-TEST program of the SPSS computer package was used to perform the analyses.

8. Determination of fit to original assessment test. In employing the Conventional and Rasch approaches to item analysis in this study only statistical information relating to items was considered. A test item was selected/rejected from the subtest being built solely on the basis of conventional or Rasch criteria applied to its item indices. A purely statistical approach to test construction is not followed in practice. Items with low indices may still relate to the content domain being measured by the test (Cronbach, 1971). In actual practice, items with inappropriate indices are improved upon where possible and carried into the final version of a test if it is felt that they measure an important area of the content domain. This is especially true when the number of items being piloted is small.

Content validity of the test items was a primary consideration in the construction of the Grade 4 Science Assessment Test. Because the number of items piloted was small, most items displaying poor item indices were refined and carried into the final version of the assessment test. To determine the fit of the subtests constructed in this study to the original assessment test the percentage of items on each subtest which also were used in the Concepts and Processes goal areas of the final assessment test was computed. Table 5 lists these items. The numbers used for the items are the ones assigned to them in the present study.

Table 5

Items Used in the Final Version of the Grade 4 Assessment Test

Understanding Concepts (26 items)
6, 7, 10, 12, 13, 19, 20, 21, 24, 27, 28, 30, 34, 38, 40, 42, 45, 46, 51, 52, 53, 54, 55, 58, 67, 72
Skill in the Processes of Science (28 items)
Classifying (5 items): 9, 11, 44, 49, 65
Observing (5 items): 16, 17, 23, 35, 60
Communicating (6 items): 14, 31, 32, 59, 70, 71
Measuring (7 items): 8, 18, 29, 37, 41, 48, 63
Combined Skills (5 items): 22, 33, 36, 56, 62

CHAPTER IV

RESULTS

The format for Chapter IV follows the sequence of steps outlined in Chapter III. A description of the final sample used in the study is provided. A brief summary of the seven different item analysis procedures carried out on the item pool follows. The results of comparisons made between the subtests resulting from the seven different approaches are then presented.

Final Sample

In order to eliminate the possible effects of speededness interfering with the Rasch item analysis, the decision rules for elimination of test booklets described in Chapter III were applied. This procedure led to the elimination of 11 booklets. Consequently, the final sample size was 516, with 246 grade four girls and 270 grade four boys.

The subjects came from three adjacent school districts located in the South Okanagan. The two districts with the smallest school populations were semirural with schools located in small towns. The district with the largest school population was a small urban center.

Agriculture is an important facet of the economy in all three districts. In the small urban center, the service industries and trade are important contributors to the economy. This city has become the leading center for manufacture of mobile homes in the Province. All three districts are important retirement and tourist areas. Logging and mining are small but still important contributors to the economy of the region.

Population growth in the three districts ranged from 15.1% to 21.8% over the span of years 1971 to 1976. English is the mother tongue spoken by the majority of the people in the area. Less than one percent of the population lists their mother tongue as being Native Indian, Chinese/Japanese, or Indo-Pakistani.

Item Analysis and Test Construction

For reasons outlined in Chapter III, it was decided to restrict the analyses to the Concepts and Processes item pools. Accordingly, the unidimensionality of the two item pools was tested using Cronbach's coefficient alpha and factor analysis as detailed in the previous chapter. It was decided to perform a series of seven item analyses--one conventional and six applying the Rasch approach--to the Concepts and Processes item pools. To facilitate subsequent presentation of the results obtained by applying the test construction criteria outlined in the previous chapter, the codes summarized in Table 6 will be used. Application of the criteria in the seven different approaches is summarized in Table 7. Raw data from the LERTAP and BICAL computer programs are presented for each item in Appendix C.

Table 6

Criteria Used in the Item Analyses

Code	Conventional Criteria
a	The difficulty index, $\underline{p}$, must lie within the range $.20 \leq \underline{p} \leq .80$.
b	The corrected item-subtest point-biserial correlation coefficient must be equal to or greater than .20.
c	All distractors must have uncorrected item-subtest point-biserial correlation coefficients less than zero.
d	All distractors must have a nonzero difficulty index.
	Rasch Criteria
MS	The mean square fit must be less than one standard deviation above unity.
DI	The discrimination index must be within one standard deviation of the mean item discrimination.
DF	The difficulty index must be within one standard deviation of the mean item difficulty.

Subtest Comparisons

1. Percentage of items rejected by each approach. The percentage of items rejected by each of the item analysis approaches is summarized in Table 8. The percentage is given for each Concepts and Processes subtest. The number of items in each item pool, k, is indicated.

2. Percentage overlap of rejected items. The percentage overlap of items commonly rejected by two different approaches was computed for each possible pair of Concepts subtests and each possible pair of Processes subtests. These percentages are summarized in Tables 9 and 10.

3. Rasch elimination of items not meeting conventional criteria. The sequence number of items rejected by application of the Conventional Approach is shown in Table 11. The conventional criteria which determined the rejection of the items are coded as in Table 6. By using the information in Table 11, it is possible to determine the percentage of these items which were also

Table 7

Item Analysis Criteria

Approach	Correction for Guessing	Criteria
Conventional	none	a, b, c, d
Rasch I	Panchapakesan	MS, DI, DF
Rasch II	Panchapakesan	MS, DI
Rasch III	Panchapakesan	MS
Rasch IV	random	MS, DI, DF
Rasch V	random	MS, DI
Rasch VI	random	MS

Table 8

Percentage of Items Rejected

Approach	Percentage of Items Rejected	
	Concepts ($\underline{k}$ = 30)	Processes ($\underline{k}$ = 32)
Conventional	40	28
Rasch I	53	41
Rasch II	30	22
Rasch III	10	6
Rasch IV	53	44
Rasch V	37	16
Rasch VI	13	3

rejected in each of the Rasch approaches. The percentage effectiveness of the Rasch approaches in identifying these items is summarized in Tables 12 and 13. Percentage Rasch efficiencies in rejecting items having specific conventional problems or combinations of several problems is shown as well as percentage efficiency in rejecting all items in Concepts and Processes subtests which display conventional criteria problems. The number of items displaying each conventional problem is shown in brackets.

4. Rasch specific items. For each Rasch approach, the percentage of items rejected solely on the basis of Rasch criteria was determined. These percentages are shown for each Concepts and Processes subtest in Table 14.

Table 9

Percentage Overlap of Rejected Items

Concepts Subtest

Approach	I	II	III	IV	V	VI
Conventional	56	40	25	47	35	33
Rasch I		56	19	68	42	25
Rasch II			33	32	43	30
Rasch III				19	27	75
Rasch IV					69	25
Rasch V						36
Rasch VI						

Table 10

Percentage Overlap of Rejected Items

Processes Subtest

Approach	I	II	III	IV	V	VI
Conventional	57	23	22	44	8	11
Rasch I		54	15	69	20	8
Rasch II			29	31	33	14
Rasch III				7	17	50
Rasch IV					36	7
Rasch V						20
Rasch VI						

Table 11

Items Rejected--Conventional Approach

Criteria	Sequence Numbers of Items Rejected	
	Concepts	Processes
a	6, 21, 38, 54	11, 14, 17, 22, 35, 49, 71
b	46, 58, 67	73
c	10, 42	--
b, c	30	--
a, b, c	24, 53	--
a, c, d	--	16
Total	12 items	9 items

Table 12

Rasch Approaches--Concepts Subtest

Percentage of Items Rejected

According to Conventional Criteria

Approach	Criteria					
	a(4)	b(3)	c(2)	b,c(1)	a,b,c(2)	Total(12)
Rasch I	100	67	50	100	100	83
Rasch II	25	33	50	100	100	50
Rasch III	0	33	0	0	100	25
Rasch IV	100	100	0	0	100	75
Rasch V	25	100	0	0	100	50
Rasch VI	0	67	0	0	100	33

Table 13

Rasch Approaches--Processes Subtest

Percentage of Items Rejected

According to Conventional Criteria

Approach	Criteria			
	a(7)	b(1)	a, c, d(1)	Total (9)
Rasch I	86	100	100	89
Rasch II	29	100	0	33
Rasch III	14	100	0	22
Rasch IV	71	100	100	78
Rasch V	0	100	0	11
Rasch VI	0	100	0	11

Table 14

Percentage of Rasch-Specific Items

Approach	Percentage of Rasch-Specific Items	
	Concepts	Processes
Rasch I	38	38
Rasch II	33	57
Rasch III	0	0
Rasch IV	44	50
Rasch V	45	80
Rasch VI	0	0

5. Subtest characteristics. Comparisons of the Concepts and Processes subtests on the basis of number of items, test means, test standard deviations, range of scores, Hoyt reliability estimates, and standard errors of measurement are summarized in Tables 15 and 16.

6. Correlations among subjects' scores on subtests. The correlations among subjects' scores on each possible pair of Concepts subtests and on each possible pair of Processes subtests were obtained. The scores used were the subjects' arcsine transformed proportional scores on each subtest. The results of these analyses are summarized in Tables 17 and 18.

7. Comparisons of subtest means. Correlated t-tests were performed to determine whether significant differences existed between the means of all possible pairs of Concepts subtests and all possible pairs of Processes subtests. The level of probability for the rejection of the null hypothesis of no significant differences between means was set at .05. The results of these analyses are summarized in Tables 19 and 20. If a significant difference was found between means, the tables indicate the approach which yielded the higher mean. An asterisk indicates no significant difference found between the two means compared.

8. Determination of fit to original assessment test. The fate of each item in each of the seven approaches to item analysis is summarized in Tables 21 and 22. Rejected items are indicated with an 'X'. The sequence number of the item is the one used in this study. An item number with an asterisk indicates that the item appeared in the final version of the Grade 4 Science Assessment Test. In an attempt to assess the fit of each subtest constructed in this study to the original assessment test, the percentage of items selected by each approach which also were contained in the Concepts and Processes categories of the final assessment test was determined. The results of

this analysis are summarized in Table 23.

A discussion of the results presented in this chapter and conclusions based upon them is presented in Chapter V.

Table 15

Test Characteristics--Concepts Subtest

Approach	Number of Items	Test Mean (%)	Test S.D. (%)	Range of Scores (%)	Hoyt Reliability	S.E.M.
Conventional	18	54.44	20.00	0.00 - 100.00	0.73	1.83
Rasch I	14	48.64	19.64	0.00 - 100.00	0.61	1.65
Rasch II	21	57.10	17.57	4.76 - 100.00	0.71	1.93
Rasch III	27	55.74	17.56	7.41 - 96.30	0.77	2.21
Rasch IV	14	50.64	19.71	0.00 - 92.86	0.61	1.67
Rasch V	19	59.11	17.79	5.26 - 94.74	0.68	1.85
Rasch VI	26	56.69	17.85	3.85 - 96.15	0.77	2.16

Table 16

Test Characteristics--Processes Subtest

Approach	Number of Items	Test Mean (%)	Test S.D. (%)	Range of Scores (%)	Hoyt Reliability	S.E.M.
Conventional	23	55.09	20.74	4.53 - 100.00	0.81	2.05
Rasch I	19	57.32	21.47	0.00 - 100.00	0.78	1.85
Rasch II	25	63.16	18.64	8.00 - 100.00	0.81	2.01
Rasch III	30	63.45	17.93	10.00 - 100.00	0.83	2.19
Rasch IV	18	61.11	21.22	0.00 - 100.00	0.78	1.76
Rasch V	27	64.81	17.74	7.41 - 100.00	0.81	2.03
Rasch VI	31	64.16	17.87	9.68 - 100.00	0.83	2.22

Table 17

Correlations Between Subjects' Scores

Concepts Subject

Approach	I	II	III	IV	V	VI
Conventional	.92	.94	.96	.90	.92	.96
Rasch I		.95	.91	.92	.90	.92
Rasch II			.97	.88	.93	.96
Rasch III				.91	.95	.99
Rasch IV					.96	.92
Rasch V						.96
Rasch IV						

Table 18

Correlations Between Subjects' Scores

Processes Subtest

Approach	I	II	III	IV	V	VI
Conventional	.97	.97	.99	.95	.96	.98
Rasch I		.98	.97	.97	.96	.97
Rasch II			.99	.96	.98	.98
Rasch III				.95	.98	.99
Rasch IV					.97	.97
Rasch V						.98
Rasch VI						

Table 19

Correlated t-tests--Concepts Subtests

Approach	I	II	III	IV	V	VI
Conventional	Con.	II	III	IV	V	VI
Rasch I		II	III	IV	V	VI
Rasch II			II	II	V	*
Rasch III				III	V	VI
Rasch IV					V	VI
Rasch V						V
Rasch VI						

Table 20

Correlated t-tests--Processes Subtest

Approach	I	II	III	IV	V	VI
Conventional	I	II	III	IV	V	VI
Rasch I		II	III	IV	V	VI
Rasch II			*	II	V	VI
Rasch III				III	V	VI
Rasch IV					V	VI
Rasch V						V
Rasch VI						

Table 21

Items Rejected--Concepts Subtests

Item Number	Approaches						
	Con.	I	II	III	IV	V	VI
1							
3		X	X		X	X	
4		X	X		X	X	
5					X	X	
6*	X	X			X		
7*							
10*	X						
12*					X	X	
13*							
19*							
20*							
21*	X	X			X		
24*	X	X	X	X	X	X	X
27*		X	X				
28*							
30*	X	X	X				
34*		X			X		
38*	X	X	X		X	X	
40*							
42*	X	X	X				
45*							
46*	X	X	X	X	X	X	X
51*							
52*		X			X	X	
53*	X	X	X	X	X	X	X
54*	X	X			X		
55*							
58*	X				X	X	
67*	X	X			X	X	X
72*		X			X		

Note: Items marked with an asterisk were those included in the original assessment test.

Table 22

Items Rejected--Processes Subtest

Item Number	Approach						
	Con.	I	II	III	IV	V	VI
8*							
9*		X	X		X	X	
11*	X	X			X		
14*	X						
16*	X	X			X		
17*	X	X	X		X		
18*							
22*	X	X	X	X			
23*							
26		X			X		
29*							
31*		X	X		X	X	
32*					X		
33*							
35*	X	X			X		
36*							
37*							
41*							
44*					X	X	
48*							
49*	X	X			X		
56*					X	X	
59*							
60*		X	X				
62*							
63*		X	X		X		
65*							
70*							
71*	X	X			X		
73	X	X	X	X	X	X	X
74							
76							

Note: Items marked with an asterisk were those included in the original assessment test.

Table 23

Percentage of Items in the Final Assessment Test Retained by Each Approach

Category	Approach						
	Con.	I	II	III	IV	V	VI
Concepts	54	46	73	88	50	69	85
Processes	71	61	79	96	57	86	100

CHAPTER V

SUMMARY AND CONCLUSIONS

Summary

Purpose and Procedure

The purpose of this study was to compare the results of applying conventional and Rasch item analysis procedures to a pool of test items. The item pool used was a 76-item modified version of the pilot tests used in constructing the British Columbia Grade 4 Science Assessment Test administered in 1978. The item pool was administered to 527 grade 4 students in 15 schools located in three adjacent school districts in the South Okanagan Valley. To control for possible effects of speededness, subjects' test booklets were eliminated from the item analysis using the two criteria outlined in Chapter III. Eleven booklets were dropped, leaving 516 booklets for item analysis.

Preliminary analyses indicated problems as regards Rasch analysis of items in short subtests of four or five items. For this reason, it was decided to restrict item analysis to the two largest subtests in the item pool. These were the Concepts and Processes subtests consisting of 30 and 32 items, respectively. To determine the unidimensionality of each of the two subtests, coefficient alpha internal consistency indices as well as factor analysis data (Table 3) were considered. Coefficient alpha information indicated strong subtest homogeneity. For the factor analyses, use of the Kaiser-Guttman criterion indicated a greater number of factors to be retained than did the scree test. In all cases, more than one common factor was found on which there were salient loadings. Inspection of clusters of items with salient pattern coefficients failed, however, to provide conceptually clear definitions of possible different traits being measured by the subtests. It was felt that use of dichotomous test data and tetrachoric correlations

among test items led to the large number of factors in each subtest. It was decided that the subtests were essentially unidimensional and that application of the Rasch model would be appropriate.

Four conventional criteria relating to item-difficulty, item-subtest point-biserial correlations, distractor-subtest point-biserial correlations, and distractor difficulty were applied in the Conventional Approach. The item-subtest point-biserial correlations were corrected for spuriousness using Guilford's formula. Three Rasch item analysis criteria relating to mean square fit, item discrimination, and item difficulty were used in Rasch Approach I. For Rasch Approach II, the item difficulty criterion was omitted, and in Rasch Approach III the only criterion used was mean square fit. The criteria used in Rasch Approaches IV, V, and VI paralleled those used in Rasch Approaches I, II, and III, respectively. For Rasch Approaches I, II, and III, Panchapakesan's correction for guessing formula was used to eliminate subjects with lower scores from the calibration sample. For Rasch Approaches IV, V, and VI, the random guessing level formula was employed to determine the minimum scores of subjects allowed in the calibration sample. The item analysis criteria used in the Conventional Approach and the six Rasch Approaches are summarized in Tables 6 and 7. The LERTAP and BICAL computer programs were used in the analysis of test items for the Conventional and Rasch Approaches, respectively.

A series of eight analyses were carried out to compare the one Conventional and six Rasch Approaches to test item analysis. Four of the analyses were focussed at the item level. The percentage of items rejected by each approach as well as the percentage overlap of rejected items among the different approaches was determined. The efficacy of the Rasch Approaches in identifying items illustrating problems as detected by the Conventional Approach

was investigated. For each of the Rasch Approaches, the percentage of items rejected solely on the basis of the Rasch criteria was determined.

Three comparisons were aimed at the subtests considered as entities. These analyses involved--(a) comparisons of numbers of items in each subtest, subtest means, standard deviations, and range of scores, as well as Hoyt estimates of reliability and subtest standard errors of measurement; (b) correlations of subjects' scores among all Concepts subtests and among all Processes subtests, and (c) correlated t-tests among all possible pairs of Concepts subtest means and all possible pairs of Processes subtest means.

The eighth analysis involved a comparison of the fit of each of the Concepts and Processes subtests resulting from the seven different approaches to the original assessment test. In this investigation, the items in the Concepts and Processes categories of the final version of the Grade 4 Science Assessment Test as determined by the Contract Team (British Columbia Ministry of Education, 1979) were used as the basis for roughly determining content validity of the tests produced by the approaches employed in this study.

Discussion of Results

1. Percentage of items rejected by each approach. An inspection of the results summarized in Table 8 indicates that the most stringent set of Rasch criteria applied in Rasch Approaches I and IV led to the rejection of the highest percentage of items. Application of the conventional criteria led to the next highest percentage rejection of items, followed by Rasch Approaches II and V. Rasch Approaches III and VI, employing only the mean square fit criterion, led to the lowest percentage rejection of items. In all approaches, a higher percentage of Concepts items as compared to Processes items was rejected.

For the Concepts subtest, Rasch Approaches V and VI, employing the random guessing level formula, led to a slightly higher percentage of items

rejected than did the parallel Rasch Approaches II and III, using Panchapakesan's formula to determine the make-up of the calibration sample. For the Processes subtest, however, the reverse is true, with Rasch Approaches II and III rejecting a somewhat higher percentage of items than the parallel Rasch Approaches V and VI.

In Conventional and Rasch item analysis, the percentage of items rejected appears to be strongly related to the stringency of the criteria employed. For the Rasch Approaches, there appears to be no appreciable difference in the percentage of items rejected by approaches using Panchapakesan's formula and parallel approaches employing the random guessing level formula.

2. Percentage overlap of rejected items. Inspection of Tables 9 and 10 indicates that the highest percentage overlap of items rejected by the Conventional Approach and the Rasch Approaches was with Rasch Approaches I and IV, in which the most stringent Rasch criteria were applied. In general, the percentage overlap of items rejected by the Conventional and Rasch Approaches becomes reduced as the Rasch criteria used in item rejection become less stringent. There is, in general, a somewhat higher percentage overlap of rejected items between the Conventional Approach and Rasch Approaches using Panchapakesan's formula for determining the calibration sample compared to Rasch Approaches employing the random guessing level formula. Figures extracted from Tables 9 and 10 illustrate the above trends in Table 24.

When the Rasch Approaches are considered alone, the highest percentage overlap of rejected items occurs between approaches using Panchapakesan's formula and their parallel approaches employing the random guessing level formula--Rasch I and IV; Rasch II and V; Rasch III and VI. Thus, in determining similarity between Rasch approaches, the number of item rejection criteria in the approaches appeared to override the effect of the guessing

Table 24

Percentage Overlap of Rejected Items

Conventional and Rasch Approaches

Approach	Conventional	
	Concepts	Processes
Rasch I	56	57
Rasch II	40	23
Rasch III	25	22
Rasch IV	47	44
Rasch V	35	8
Rasch VI	33	11

formula applied in determining the calibration sample. These trends are summarized in Table 25.

Table 25

Percentage Overlap of Rejected Items

Parallel Rasch Approaches

(Concepts)	IV	V	VI
Rasch I	68		
Rasch II		43	
Rasch III			75
(Processes)			
Rasch I	69		
Rasch II		33	
Rasch III			50

There is a high percentage overlap of rejected items between Approaches I and II in which Panchapakesan's formula was applied and in general between Approaches IV and V in which the random guessing level formula was used. The

percentage overlap of rejected items between Approaches II and III, and between Approaches V and VI is moderate. The lowest percentage overlap of rejected items occurs when Approaches I and III, and when Approaches IV and VI are compared. When the following comparisons across different methods of determining the calibration sample are made--Approaches I and V; Approaches I and VI; Approaches II and VI--the percentage overlap of rejected items tends to drop as compared to parallel comparisons within the same methods of determining the calibration sample--Approaches I and II; Approaches I and III; Approaches II and III. These trends are summarized in Table 26.

Table 26

Percentage Overlap of Rejected Items

Rasch Approaches Within (a) and Across (b)

Methods of Determining the Calibration Sample

(Concepts)	II	III	V	VI
Rasch I	56[a]	19[a]	42[b]	25[b]
Rasch II		33[a]		30[b]
Rasch IV			69[a]	25[a]
Rasch V				36[a]
(Processes)				
Rasch I	54[a]	15[a]	20[b]	8[b]
Rasch II		29[a]		14[b]
Rasch IV			36[a]	7[a]
Rasch V				20[a]

Two factors appear to influence the percentage overlap of rejected items among the Rasch approaches. A strong factor is the number of criteria involved. Approaches employing parallel criteria but different guessing level formulas generally illustrate a high percentage overlap of rejected items. Also, comparisons involving one criterion/two criteria approaches and two criteria/three criteria approaches rejected a larger percentage of items in common

than did comparisons involving one criterion/three criteria approaches. Given the number-of-criteria factor, comparisons involving one criterion/two criteria approaches and two criteria/three criteria approaches rejected a larger percentage of items in common when the same guessing formula was applied to determine the calibration sample than when parallel comparisons across different types of guessing formulas were made. The number-of-criteria relationship between approaches, however, appears to be much stronger than effects caused by different guessing level formulas.

3. Rasch elimination of items not meeting conventional criteria. The specific items rejected by the Conventional Approach are summarized in Table 11. The percentages of these items also rejected by the six Rasch approaches were computed and are summarized in Tables 12 and 13.

The overall identification of items displaying problems related to the conventional criteria was highest for the three criteria Rasch Approaches I and IV. The percentages progressively dropped for the two criteria Approaches II and V, and the one criterion Approaches III and VI. Particularly for the Processes subtest, approaches employing the random guessing level formula identified fewer items having conventional problems than did parallel approaches employing Panchapakesan's formula.

As regards specific problems with items, it seemed clear that removal of the Rasch item difficulty criterion lowered the percentage identification of items having a p-value outside the range .20 to .80 (conventional criterion a). In general, the approaches employing only the mean square fit criterion were insensitive as regards the identification of items with too high or too low p-values. Particularly for the Processes subtest, approaches employing Panchapakesan's formula seemed to be somewhat more effective in identifying items with p-value problems.

A clear trend was not discernable as regards Rasch identification of

items having item-subtest point-biserial correlation coefficients less than .20 (conventional criterion b). In the case of the Processes subtest, all approaches were effective in identifying the single item displaying this problem. In the case of the Concepts subtest, it could not be determined whether removal of the item difficulty criterion or removal of the discrimination criterion had a larger effect in reducing the power of the analyses to identify items with point-biserial problems. Approaches using the random guessing level formula appeared to be somewhat more effective in identifying point-biserial problems than did approaches using Panchapakesan's formula.

The Rasch approaches seemed rather insensitive as regards the identification of items displaying problems with distractors (conventional criteria c and d). For the Concepts subtest, only Rasch Approaches I and II identified items having distractors with a point-biserial coefficient greater that zero. All approaches rejected some items having distractor problems when these occured in conjunction with p-value and/or item-subtest point-biserial problems (conventional criteria combinations b, c; a, b, c, and a, c, d). In general, approaches employing Panchapakesan's formula were somewhat more effective in identifying items with distractor problems.

4. Rasch specific items. Percentages of items rejected by the different Rasch approaches solely on the basis of Rasch criteria as compared to the total number of items rejected by each approach are summarized in Table 14. The figures indicate that all items rejected by the one criterion Approaches III and VI displayed conventional problems as well. In general, the two criteria Approaches II and V rejected the largest percentages of items not displaying conventional problems except in the case of the Rasch II Approach applied to the Concepts subtest. Generally, a somewhat smaller percentage of items rejected specifically by Rasch criteria was indicated for the three criteria Approaches I and IV except in the case of Rasch Approach I applied

to the concepts subtest. Approaches using the random guessing level formula appeared to reject more Rasch-specific items than did approaches employing Panchapakesan's formula.

5. Subtest characteristics. Subtest characteristics are summarized in Tables 15 and 16. The lengths of the Concepts subtests ranged from 14 to 27 items, while the lengths of the Processes subtests ranged from 18 to 31 items. As could be expected, approaches with less stringent criteria yielded the longest subtests. Subtests yielded by the one criterion Approaches III and VI were longest followed by subtests yielded by the two criteria Approaches II and V. These were followed in length by the subtests built using the Conventional Approach, and finally by subtests constructed using the three criteria Rasch Approaches I and IV.

The range in subtest means was not large--Concepts 48.64% to 59.11%; Processes 55.09% to 64.81%. Making comparisons among Rasch approaches in which the same guessing level formula was applied, the two criteria approaches yielded subtests with the highest means. These were followed by the means of subtests constructed using the one criterion approach, subtests constructed using the Conventional Approach, and finally subtests built using the three criteria approach. In the case of the Conventional Approach, the low subtest means resulted from the fact that, of the 14 items with a p-value outside the range of .20 to .80, 12 were easy items with a p-value greater than .80. Elimination of these items tended to lower subtest means. Rasch Approaches I and IV, using an item difficulty criterion, tended to eliminate some of these easy items as well, while Approaches II, III, V, and VI eliminated fewer or no items with conventional difficulty related problems. This resulted in higher subtest means for the latter Rasch approaches. This effect is also evident in inspecting the lower ranges of scores obtained on subtests

built using Rasch Approaches I and IV, which eliminated some of the very easy items. Subtest means were generally somewhat higher for subtests constructed using Rasch approaches in which the random guessing level formula was used compared to subtests built using parallel item rejection criteria but applying Panchapakesan's formula to determine the make-up of the calibration sample. This trend is illustrated in Figure 3.

In general, the Hoyt internal consistency estimates for tests constructed using Rasch approaches was greatest for Approaches III and VI. These subtests had the greatest number of items. Then followed internal consistency reliabilities obtained for subtests constructed using the Conventional Approach, followed by coefficients for subtests built using Rasch Approaches II and V, althougn subtests constructed using the Conventional Approach had somewhat fewer items than did subtests constructed using Approaches II and V. The subtests with the least number of items, constructed using Rasch Approaches I and IV, had the lowest Hoyt reliability indices. There appeared to be no major differences in Hoyt reliability estimates for subtests constructed using the random guessing level formula and those constructed using Panchapakesan's formula.

It was evident that the number of items in the subtests resulting from each approach was having an effect on the Hoyt reliability estimates. In an effort to control for this factor, stepped-up reliability estimates for subtests of equal length produced by each approach were calculated. The formula used was the Spearman-Brown prophecy formula (Magnusson, 1967):

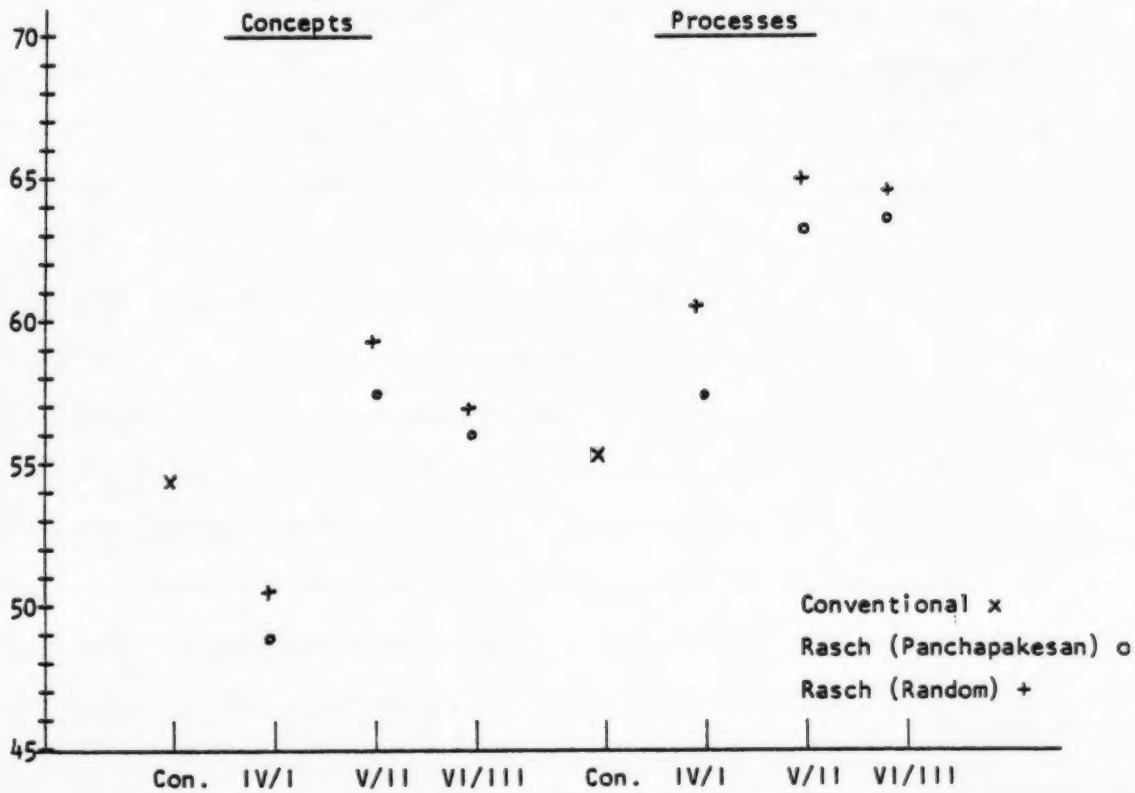
$$\underline{r}_{\underline{tt}} = \frac{\underline{N}'\underline{r}}{1 + (\underline{N}' - 1)\underline{r}}$$

where $\underline{N}'$ = the number of times the test is longer or shorter than the original test, and

$\underline{r}$ = the reliability of the test which is being lengthened or shortened.

Figure 3

Subtest Means (%)



The reliabilities of the Concepts subtests were stepped-up to a test of 30 items in length--the maximum number of items in the Concepts item pool. The reliability of the Processes subtests were stepped-up to a test of 32 items in length. The results are summarized in Table 27. The results emphasize the lack of any real differences among the reliabilities of the subtests constructed by applying the different approaches.

Table 27

Stepped-up Hoyt Reliability Estimates

Approach	Subtest	
	Concepts	Processes
Conventional	.82	.86
Rasch I	.77	.86
Rasch II	.78	.85
Rasch III	.79	.84
Rasch IV	.77	.86
Rasch V	.77	.83
Rasch VI	.79	.83

Lord (1959) and Cureton et al, (1973) have shown empirically that the standard error of measurement, SEM, of a test is directly proportional to $\sqrt{k}$, the square root of the number of items on a test. An attempt was made to determine the nature of this relationship for the subtests constructed using the approaches in this study. Subtest SEM's were correlated with the corresponding $\sqrt{k}$'s, and the linear regression of the SEM's on the $\sqrt{k}$'s was also obtained. Analyses were performed for all 14 subtests (two conventional; 12 Rasch), for the 12 Rasch subtests, for the six Rasch subtests constructed using Panchapakesan's formula, and for the six Rasch subtests constructed using approaches employing the random guessing level formula. The results of these analyses are summarized in Table 28.

Table 28

Relationships Between Subtest SEM's and $\sqrt{k}$'s

Analysis	Correlation	Slope
All subtests	.965	.314
All Rasch subtests	.971	.320
Rasch subtests (Panchapakesan)	.976	.329
Rasch subtests (random guessing)	.968	.311

A high correlation between subtest SEM's and the square roots of the number of items on the subtest is evident. The values of the regression coefficients are considerably lower than the findings of Lord, or Cureton et al. In any case, subtest SEM's appear to be solely affected by the number of items in the subtests and not by the different approaches used in their construction.

6. Correlations among subjects' scores on subtests. The correlations among subjects' scores on each possible pair of subtests are shown in Tables 17 and 18. As can be seen, regardless of the item analysis approach used, subtest scores are highly correlated. Subtests built using the different approaches tend to order subjects in much the same fashion.

7. Comparison of subtest means. The means of the subtests produced by the different approaches have been discussed under section 5 above. The results of correlated t-tests comparing each possible pair of subtest means are presented in Tables 19 and 20. Tables 15 and 16 indicate that the differences between many of the subtest means are not great. The results of the t-tests, however, show that only two subtest means, those for Rasch Approaches II and VI (Concepts), and those for Rasch Approaches II and III (Processes) do not differ significantly at the $\alpha = .05$ level. It is felt that the large n of 516 involved in the t-tests, as well as the high correlations between

subjects' scores on the subtests increased the power of the t-tests to the extent that significant differences between subtest means were indicated in the majority of the comparisons.

8. Determination of fit to original assessment test. Most of the items used in the pilot testing for the Grade 4 Science Assessment Test were retained in the final version of that test. Since the final version of the assessment test was the standard by which the content sampling of the Concepts and Processes subtests constructed in this study was judged, approaches using less stringent criteria and thus rejecting fewer items sampled the content domain most effectively. As shown in Table 23, subtests constructed using Rasch Approaches III and VI retained the highest percentage of items used in the original test. These were followed by the subtests built using the two criteria Rasch Approaches II and V. The Conventional Approach and Rasch Approaches I and IV, employing the most stringent sets of criteria yielded subtests with the lowest percentages of items used in the original test. Obvious differences in sampling the content domain for subtests built by Rasch approaches using the random guessing level formula as compared to those using Panchapakesan's formula were not discernable.

Conclusions

The one Conventional and six Rasch approaches utilized in this study produced quite different subtests as regards the specific items selected. The numbers of items in the subtests and the subtest characteristics, however, were more a function of the stringency of the criteria applied, rather than of the particular approach, Conventional or Rasch, used to build the subtests. Indeed, the highest percentage overlap of items rejected for the Conventional Approach was with Rasch Approaches I and IV. These three approaches applied the strictest criteria for item rejection.

The number of items in each subtest, in turn, affected the reliability

of the instruments. When the Spearman-Brown prophecy formula was applied in order to compare the Hoyt internal consistency estimates of the subtests on a par with each other, no apparent differences were found in the adjusted coefficients. Test SEM's and the square roots of the number of items on the subtests were found to be highly correlated. Regression of subtest SEM's on the respective $\sqrt{k}$'s provided fairly consistent results for all subtests, all Rasch subtests, Rasch subtests built using Panchapakesan's formula, and Rasch subtests built using the random guessing level formula.

The fit of the subtests to the original assessment test was also apparently a function of the stringency of the criteria used in the different item analysis approaches. Application of stringent criteria, whether Conventional or Rasch, produced subtests with fewer items and these shorter subtests provided a poorer sampling of the content domain than did subtests constructed using more lenient approaches.

Selection of items solely on the basis of statistical criteria, as done in this study, is rarely carried out in actual test construction tasks. When faced with a limited number of items to test given goals and objectives, as were the test builders in constructing the Grade 4 Science Assessment Test, the data provided by conventional item analysis is more helpful than item data arrived at through Rasch item analysis. Conventional item analysis criteria can be applied to identify faulty items which can then be improved upon. Such items can be rewritten and the shrinkage of items available to test the desired knowledge and skills will be minimal.

Stringent application of Rasch item analysis criteria, however, would appear to lead to a dead end. The item indices do not appear to carry the information needed for refinement of faulty items. There appears to be an analagous relationship between the conventional p-value and the Rasch item difficulty index. Identification of items having low item-subtest correla-

tions is somewhat less effective when Rasch item analysis procedures are used. Also, Rasch item analysis appears to be particularly impotent in identifying distractor problems as revealed by conventional item analysis.

Although Rasch approaches seem to produce subtests of equal reliability compared to the conventional approach, and Rasch subtests order subjects in very much the same fashion as subtests built using conventional methods, the findings of this study indicate that the conventional approach should be the method of choice especially when small numbers of items are available for pilot testing. The Rasch approach to item analysis might find effective application when large numbers of items are available for pilot testing. Use of Rasch item analysis, however, would incur the expense of rejection of valuable items which could be improved upon and used effectively.

Should the claims of specific objectivity of item and person ability parameters tendered by proponents of the Rasch model someday be proved unequivocally, the importance of the Rasch model would be greatly enhanced. At such a time, a hybrid approach to item analysis using a combination of conventional and Rasch criteria might be most potent in capitalizing on the unique qualities of both approaches. As yet, however, the Rasch approach appears to provide no unique benefits to the field of test item analysis.

While the Rasch model fails as a creative tool in test construction, it is possible that it could be employed as a confirmatory tool once tests have been constructed using conventional item analysis (Rogers, 1981). The difficulties in factor analyzing dichotomous test data are well known (Gorsuch, 1974). Factor analysis of test items usually leads to a spuriously high number of factors even though other evidence indicates that the test is in actuality unidimensional. Application of the Rasch model in determining whether a group of items measures a single trait might be less problematical

than use of factor analysis for the same purpose.

Limitations of This Study

This study did not address itself to the claims of sample-free item indices and measures of person ability associated with the Rasch model. The model was treated as an item analysis approach to be used in a straightforward test construction task. The outcomes of its application were compared with results obtained using conventional item analysis.

In interpreting the results of this study, it must be borne in mind that the test items used represented an already highly refined collection of test items. Many of them had been obtained from previously constructed tests by the Contract Team. All had undergone further refinement in the steps used to construct the Grade 4 Science Assessment Test.

The sample of students used in this study was essentially one of convenience. The subjects came from semirural and small urban areas. Generalization of the findings to other populations of interest, for instance those living in large urban areas where English may often be the second language spoken in the home, must be viewed with caution.

Directions for Future Research

As mentioned above, the items used in this study were highly refined. It is possible that comparison of Rasch and conventional item analysis approaches, applied to a pool of newly constructed test items, would produce results more clearly pointing to the superiority of one approach over the other.

The Rasch item analysis approaches were deficient as regards the identification of items having conventional problems, especially as regards distractor analysis. Research directed towards exploiting the possible information carried in Rasch item indices to identify items having specific problems would be valuable in that the Rasch approach could be applied in item analysis

with a view towards item improvement rather than rejection.

As mentioned in the review of the literature in Chapter II, the two-parameter and three-parameter latent trait models are also available for test item analysis. A comparison of the results obtained with these approaches as compared with those for conventional item analysis would be illuminating as regards the relative efficiency of latent trait item analysis as compared to the conventional approach.

Given the difficulties involved in using factor analysis to determine test homogeneity, a worthwhile avenue of research might be to investigate the efficacy of the Rasch model as a contending tool to be used in determining the unidimensionality of tests.

BIBLIOGRAPHY

Aleamoni, L. M., & Spencer, R.E. A comparison of biserial discrimination, point biserial discrimination, and difficulty indices in item analysis data. Educational and Psychological Measurement, 1969, 29, 353-358.

Andersen, E. B. A goodness of fit test for the Rasch model. Psychometrika, 1973, 38, 123-140.

Anderson, J., Kearney, G. E., & Everett, A. V. An evaluation of Rasch's structural model for test items. British Journal of Mathematical and Statistical Psychology, 1968, 21, 231-238.

Berk, R. A. Empirical evaluation of formulae for correction of item-total point-biserial correlations. Educational and Psychological Measurement, 1978, 38, 647-652.

Bloom, B. S., Hastings, J. T., & Madaus, G.F. Handbook of Formative and Summative Evaluation. New York: McGraw-Hill, 1971.

British Columbia Ministry of Education. Assessment Planning: B.C. Assessment Program. Victoria, B.C., 1975.

British Columbia Ministry of Education. General Report: British Columbia Science Assessment, 1978, Vol. 1. Victoria, B.C., 1979

Brown, F. G. Principles of Educational and Psychological Testing. New York: Holt, Rinehart and Winston, 1976.

Cartledge, C. McC. A comparison of equipercentile and Rasch equating methodologies (Doctoral dissertation, Northwestern University, 1976). Dissertation Abstracts International, 1976, 37, 5141A. (University Microfilms No. 76-2215).

Cohen, L. A modified logistic response model for item analysis. Unpublished manuscript, 1976.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L. J. Essentials of Psycholotical Testing. (3rd. ed.). New York: Harper and Row, 1970.

Cronbach, L. J. Test Validation. In R. L. Thorndike (Ed.), Educational Measurement (2nd. ed.). Washington: American Council on Education, 1971.

Cureton, E. E. Corrected item-test correlations. Psychometrika, 1966, 31, 93-96.

Cureton, E. E., Cook, J. A., Fischer, R. T., Laser, S. A., Rockwell, N. J., & Simmons, J. W. Length of test and standard error of measurement. Educational and Psychological Measurement, 1973, 33, 63-68.

Dinero, T. E., & Haertel, E. Applicability of the Rasch model with varying item discriminations. Applied Psychological Measurement, 1977, 1, 581-592.

Douglass, F. M., Khavari, K. A., & Farber, P. D. A comparison of classical and latent trait item analysis procedures. Educational and Psychological Measurement, 1979, 39, 337-352.

Du Bois, P. H. Varieties of psychological test homogeneity. American Psychologist, 1970, 25, 532-536.

Ebel, R. L. Essentials of Educational Measurement. Englewood Cliffs, N. J.: Prentice-Hall, 1972a.

Ebel, R. L. Why is a longer test usually a more reliable test? Educational and Psychological Measurement. 1972b, 32, 249-253.

Educational Research Institute of British Columbia. Item Bank Development for Science 4 - 7. Edge, January - February, 1980, p. 4.

Engelhart, M. D. A comparison of several item discrimination indices. Journal of Educational Measurement, 1965, 2, 69-76.

Findley, W. G. A rationale for evaluation of item discrimination statistics. Educational and Psychological Measurement, 1956, 16, 175-180.

Forbes, D. W. The use of Rasch logistic scaling procedures in the development of short multi-level arithmetic achievement tests for public school measurement. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976. (ERIC document Reproduction Service No. ED128400).

Forster, F., Ingebo, G., & Wolmut, P. Monographs on Improving Measurement and Evaluation. Unpublished manuscript, undated.

Fryman, J. G. Application of the Rasch simple logistic model to a mathematics placement examination (Doctoral dissertation, University of Kentucky, 1976). Dissertation Abstracts International, 1976, 37, 562A. (University Microfilms No. 77-5689).

Gardner, P. L. Test length and the standard error of measurement. Journal of Educational Measurement. 1970, 7, 271-273.

Glass, G. V. & Stanley, J. C. Statistical Methods in Education and Psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Gorsuch, R. L. Factor Analysis. Philadelphia: Saunders, 1974.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. Limitations of coefficient alpha as an index of test unidimensionality. Educational and Psychological Measurement, 1977, 37, 827-838.

Guilford, J. P. The correlation of an item with a composite of the remaining items in a test. Educational and Psychological Meausrement, 1953, 13, 87-93.

Gustafsson, J-E., A review of estimation procedures for the Rasch model with an eye toward longish tests. Journal of Educational Statistics, 1980, 5, 35-64.

Hakstian, R. A., & Bay, K. S. User's Manual to Accompany the Alberta General Factor Analysis Program (AGFAP). Division of Educational Research Services, University of Alberta, 1973.

Hambelton, R. K., & Traub, R. E. Information curves and efficiency of three logistic test models. British Journal of Mathematical and Statistical Psychology, 1971, 24, 273-281.

Hambleton, R. K., Traub, R. E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 195-211.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory: models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Hashway, R. M. A comparison of tests derived using Rasch and traditional psychometric paradigms (Doctoral dissertation, Boston College, 1977). Dissertation Abstracts International, 1977, 38, 744A. (University Microfilms No. 77-17,594).

Henryssen, S. Correction of item-total correlations in item analysis. Psychometrika, 1963, 28, 211-218.

Henryssen, S. Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), Educational Measurement (2nd. ed.). Washington: American Council on Education, 1971.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Karabinus, R. A. The r-point biserial limitation. Educational and Psychological Measurement, 1975, 35, 277-282.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Lange, A., Lehmann, I. J., & Mehrens, W. A. Using item analysis to improve tests. Journal of Educational Measurement, 1967, 4, 65-68.

Lord, F. M. Tests of the same length do have the same standard error of measurement. Educational and Psychological Measurement, 1959, 19, 233-239.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Lumsden, J. The construction of unidimensional tests. Psychological Bulletin, 1961, 58, 122-131.

Magnusson, D. Test Theory. Reading, Mass.: Addison-Wesley, 1967.

Nelson, L. R. Guide to LERTAP Use and Interpretation. Dunedin, New Zealand: Department of Education, University of Otago, 1974.

Nile, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. Statistical Package for the Social Sciences (2nd. ed.). New York: McGraw-Hill, 1975.

Nunnally, J. C. Psychometric Theory. New York: McGraw-Hill, 1967.

Oosterhof, A. C. Similarity of various item discrimination indices. Journal of Educational Measurement, 1976, 13, 145-150.

O'Shea, T. J. An application of the Rasch logistic model to the assessment of change in mathematics achievement. (Doctoral dissertation, University of British Columbia, 1979).

Panchapakesan, N. The simple logistic model and mental measurement. (Unplublished doctoral dissertation, University of Chicago, 1969).

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research, 1960.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Rentz, R. R. A Short Glossary of Rasch Model Terminology. Unpublished manuscript, 1979.

Rentz, R. R., & Bashaw, W. L. The national reference scale for reading: an application of the Rasch model. Journal of Education Measurement, 1977, 14, 161-179.

Rentz, R. R., & Rentz, C. C. Does the Rasch Model Really Work? A Discussion for Practitioners (TM #67). Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, 1978.

Rogers, W. T. British Columbia learning assessment program. Studies in Educational Evaluation, 1979, 5, 39-53.

Rogers, W. T. Personal communication, August 12, 1981.

Ryan, J. P., & Hamm, D. W. Practical procedures for increasing reliability of classroom tests by using the Rasch model. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1976. (ERIC document Reproduction Service No. ED129901).

Stanley, J. C. Reliability, In R. L. Thorndike (Ed.) Educational Measurement (2nd. ed.). Washington: American Council on Education, 1971.

Stanley, J. C., & Hopkins, K. D. Educational and Psychological Measurement and Evaluation (5th. ed.). Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Tinsley, H. E. A., & Dawis, R. V. An investigation of the Rasch simple logistic model: sample free item and test calibration. Educational and Psychological Measurement, 1975, 35, 325-339.

Whitely, S. E. Models, meanings and misunderstandings: some issues in applying Rasch's theory. Journal of Educational Measurement, 1977, 14, 227-235.

Whitely, S. E., & Dawis, R. V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, 163-178.

Wolf, R. Evaluation of several formulae for correction of item-total correlations in item analysis. Journal of Educational Measurement, 1967, 4, 21-26.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977a, 14, 97-116.

Wright, B. D. Misunderstanding the Rasch model. Journal of Educational Measurement, 1977b, 14, 219-225.

Wright, B. D., & Douglas, G. A. Best procedures for sample-free item analysis. Applied Psychological Measurement, 1977a, 1, 281-295.

Wright, B. D., & Douglas, G. A. Conditional versus unconditional procedures for sample-free item analysis. Educational and Psychological Measurement, 1977b, 37, 47-60.

Wright, B. D., & Mead, R. J. BICAL: Calibrating items and scales with the Rasch model. Research Memorandum No. 23A. Statistical Laboratory, Department of Education, University of Chicago, 1978.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

APPENDIX A

TEST AND ADMINISTRATION MATERIALS

79-11-29

Dear Principal,

I wish to thank you, your grade four teachers, and the grade four students in your school for aiding me in conducting this research.

I have forwarded _____ copies of the modified version of the P.L.A.P. Grade/Level 4 Science Achievement test being used in this project. Several sets of directions for administration of the test are also included.

Please arrange, if possible, to have the test administered during the week of December 3 to 7 to the grade four students in your school. I will arrange to have the completed tests and unused copies picked up from your school's office during the following week.

Should you have any questions regarding the administering of the test, I will be pleased to answer them. Simply call me collect at Uplands Elementary School, Penticton (492-5623). Again, thank you.

Sincerely yours,

John W. Knodel

John W. Knodel
215-864 Fairview Road,
Penticton, B.C.
V2A 5Y8

MODIFIED VERSION - P.L.A.P. SCIENCE ACHIEVEMENT PILOT TEST (1977)

GRADE/YEAR FOUR

ADMINISTRATION DIRECTIONS

December 3 - 7, 1979

Thank you for your co-operation in administering this Science Pilot Test at this very busy time of year. The test is a modified version of the Year/Grade 4 P.L.A.P. Science Achievement Pilot Test which was administered in November/December 1977. The purpose of this test is NOT to measure individual, school or district achievement. Rather, item analysis data resulting from this testing will be used to build a series of tests using both the conventional and Rasch approaches to test building. The usefulness and end products of the two test construction models will then be compared.

BEFORE STARTING THE TEST

1. You may wish to examine a copy of the test. It consists of two parts: Part One - pages 1 to 16; Part Two - pages 17 - 34. Ideally, the test should be administered with a rest period between the two parts. *NOTE: If a rest period is taken, it is essential that each student retain the same booklet for both parts of the test.*

   Some of the questions asked may not seem to fit every science program because science programs in Grade/Year 4 as well as primary science programs vary. However, the test is a reflection of the original (1977) test construction committee's desire to test five basic science skill areas: Concepts, Processes, Application, Safety and Scientific Literacy. *It is essential that students attempt all questions.*

2. Before starting the test, please write the following sample question on the blackboard:

   How many months are there in a year?
   - 10 .....☐
   - 12 .....☐
   - 24 .....☐
   - 52 .....☐
   - I don't know .....☐

3. Ensure that all students have a sharpened pencil or two, a clean eraser, and a centimeter ruler (needed for question 37). You may wish to have a supply of sharpened pencils handy to take care of emergencies.

4. When the pupils are settled, distribute the booklets, telling the pupils not to open the booklets until told to do so, and to place their rulers at the top of their desks, ready for use when needed.

   Students are to answer directly in the test booklets.

over ...

STARTING THE TEST

1. Students should be informed that the purpose of this test is to provide information which will enable a researcher to build and compare different types of science tests. The students can help by working as carefully and quickly as they can, and by attempting all questions.

2. Have the students open their booklets to the Background Information section on page 1. Item 1 can be used to ensure that students obtain their original test booklet after a rest period. You may wish to have them put their name or class list number in the box. If this is done in pencil, it can be erased after the test is completed to ensure confidentiality of the student's identification. Please remember, it is important that students complete both parts of the test in the same booklet. You may wish to devise your own method of identification to ensure that each student receives his own test booklet upon resumption of the test administration after a rest period.

   Ensure that the students have time to complete items 2, 3, 4 and 5 of the Background Information. This information on age and gender may lead to valuable insights in the item analysis to be performed. Students are to place a check mark in the box to show their answers. Be sure a numeral and not the name of a day is provided for question 5.

3. Read the following instructions to the students: FOR EACH QUESTION, SELECT THE ONE BEST ANSWER AND PUT A CHECK MARK (✓) IN THE BOX OPPOSITE YOUR ANSWER. IF YOU HAVE NO IDEA OF THE CORRECT ANSWER, PLACE YOUR MARK IN THE BOX THAT SHOWS " I DON'T KNOW. "
   Have the class read the example you have written on the board and have a student suggest the correct answer. Illustrate where the check mark should go. Remind the students of the 'I don't know' option if they have no idea of the correct answer and tell them to be sure to mark only one answer space for every question. If they change an answer, they should erase their original check marks carefully.

4. Tell the students to begin with Question 1 at the bottom of page 1. Tell them to stop after they complete question 40 on page 16. If they finish earlier than other students, they may go back and check their answers, but they are not to go ahead.

5. When the second sitting resumes, have pupils turn immediately to page 17, Question 41 and conduct Part Two in the same manner as Part One.

TIME ALLOCATION

Distribution of materials; completion of background information ..... 5 to 10 minutes
Part 1 of Test ..... 30 to 35 minutes (average)*
Part 2 of Test ..... 30 to 35 minutes (average)*
* If at all possible please ensure that all, or at least most (95%) of the students in your class have time enough to complete all items. If a student is absent for part of the test, please allow him/her to make up that part, if at all possible.

AFTER TESTING

1. As soon as the test is finished have the students hand in their booklets. Stack unused tests on top of the used tests and return the stack to the principal's office.

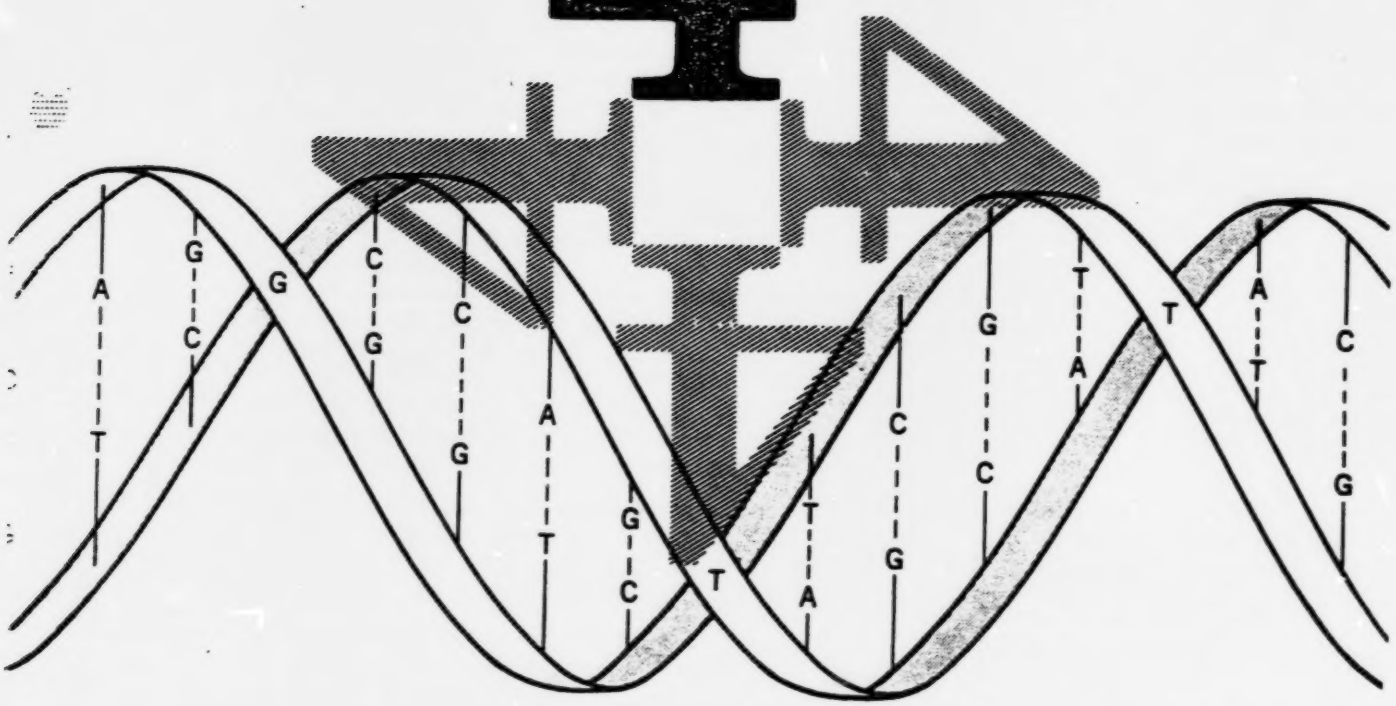
Again allow me to say how much I appreciate the participation and co-operation of you and your class in helping me carry out this research. If you have any comments or suggestions about the test or the directions for administration, I would be very pleased to hear from you.

John Knodel
215-864 Fairview Road,
Penticton, B.C.
V2A 5Y8

SCIENCE

GRADE / YEAR

4



GRADE/YEAR 4 SCIENCE

BACKGROUND INFORMATION

1. Please put your name or class number in the box.

2. **Are you a boy or a girl?**

- Boy ☐ 1
- Girl ☐ 2

3. **How old are you today?**

- 8 years of age ☐ 1
- 9 years of age ☐ 2
- 10 years of age ☐ 3
- 11 years of age ☐ 4
- more than 11 years of age ☐ 5

4. **When were you born?**

Month:
- January ☐ 01
- February ☐ 02
- March ☐ 03
- April ☐ 04
- May ☐ 05
- June ☐ 06
- July ☐ 07
- August ☐ 08
- September ☐ 09
- October ☐ 10
- November ☐ 11
- December ☐ 12

5. **Day** ☐

STOP! WAIT FOR YOUR TEACHER TO READ THE

INSTRUCTIONS

1. In order to open a can of tomato juice, Sue punched TWO holes. Why do you think she did this?

- To let the juice pour out of the can more slowly ☐ 1
- To let air get into the can as the juice was poured ☐ 2
- To let juice pour out of both holes ☐ 3
- To watch how the juice was pouring out ☐ 4
- I don't know ☐ 5

A B C D

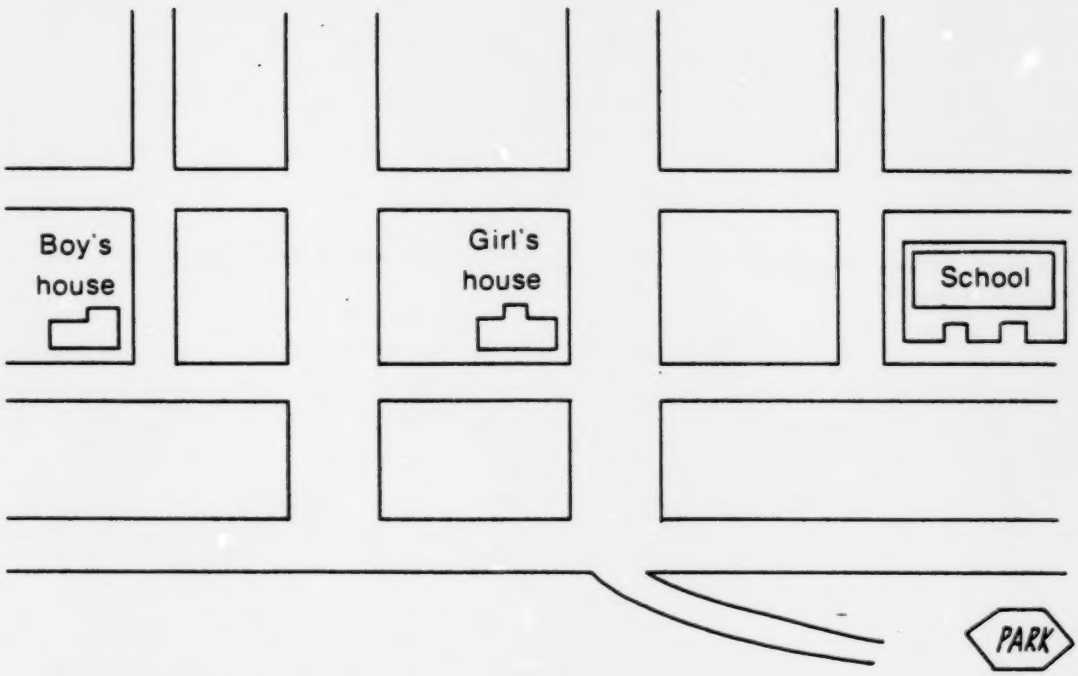
2. **Which container has a sign which means** THIS CONTAINER MAY EXPLODE?

A ☐ 1
B ☐ 2
C ☐ 3
D ☐ 4
I don't know ☐ 5

3. Most of the earth is covered by:

Fresh water ☐1
Salt water ☐2
Ice ☐3
Soil ☐4
I don't know ☐5

4. A boy and a girl both left school at the same time and arrived at their homes at the same time. They both went straight home.

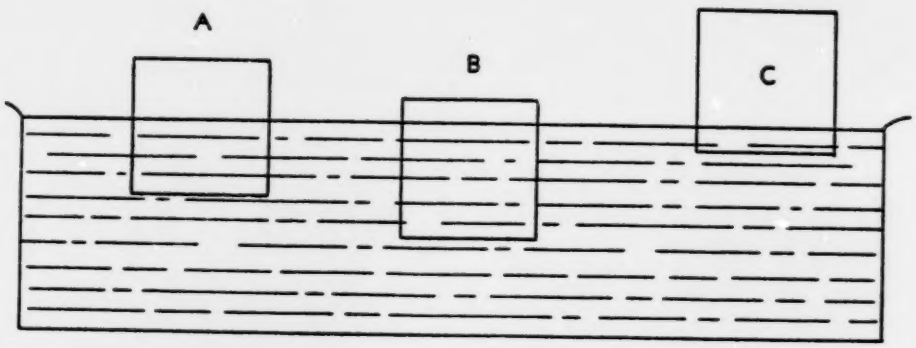


Who moved more quickly?

The girl moved more quickly ..... ☐1
The boy moved more quickly ..... ☐2
They both moved at the same speed.. ☐3
I don't know..... ☐4

5. Why does a compass needle point North?

the Earth has only one moon ..... ☐1
Earth has a magnetic field ..... ☐2
Earth reflects the Sun's light ..... ☐3
Earth's temperature is not always the same. ☐4
I don't know ..... ☐5



6. **The three solid objects shown above are the same size. If they float as shown, which one weighs the most?**

- Object A ☐ 1
- Object B ☐ 2
- Object C ☐ 3
- They all weigh the same ☐ 4
- I don't know ☐ 5

7. **Each year the Earth moves once around**

- Mars ☐ 1
- Saturn ☐ 2
- the Sun ☐ 3
- the Moon ☐ 4
- all of the other planets ☐ 5
- I don't know ☐ 6

8. **Which would be easiest to measure with a metric ruler that is 30 cm long?**

- The length of a pencil ☐ 1
- The thickness of a sheet of paper ☐ 2
- The distance from your home to the school ☐ 3
- The distance from your home to the nearest grocery store ☐ 4
- I don't know ☐ 5

All of the leaves in the first row are DICOTS

oak

strawberry

maple

rose

All of the leaves in the second row are MONOCOTS

lily

corn

plantain

bulrush

ONE **of the leaves in the third row is a monocot**

dogwood

A

pondweed

B

aspen

C

geranium

D

9. Which leaf is a monocot?

A..... ☐ 1
B..... ☐ 2
C..... ☐ 3
D..... ☐ 4
I don't know..... ☐ 5

10. **When Tom threw his rubber ball into the air, it came back to the ground because**

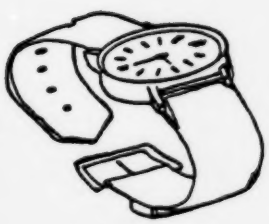
- the air pushed it back ☐ 1
- rubber always bounces back ☐ 2
- the earth pulled it back ☐ 3
- the air is very light ☐ 4
- the earth is a large magnet ☐ 5
- I don't know ☐ 6

11. **Which of these are pictures of living things?**

bicycle
A

plant
B

goldfish
C



watch
D

mountain
E

- A and B ☐ 1
- A and C ☐ 2
- B and C ☐ 3
- A, D, and E ☐ 4
- I don't know ☐ 5

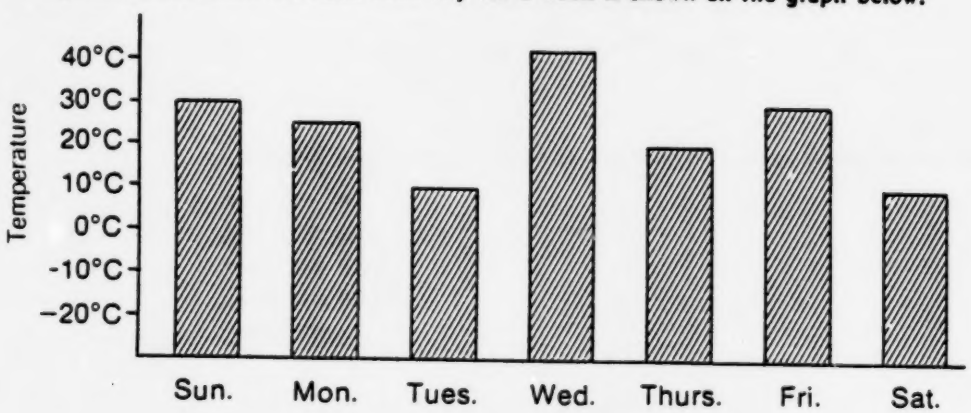
12. **The picture below shows a marble and a ruler. If the marble rolls from A to B in 2 seconds at a steady speed, how fast is it going?**

- 12 cm per 2 seconds ☐ 1
- 24 cm per second ☐ 2
- 24 mm per second ☐ 3
- 240 mL per second ☐ 4
- 12 cm per second ☐ 5
- I don't know ☐ 6

13. **In many parts of British Columbia, a noon temperature of 2°C is most likely in**

- January ☐ 1
- May ☐ 2
- June ☐ 3
- July ☐ 4
- I don't know ☐ 5

The highest temperature recorded each day for a week is shown on the graph below.



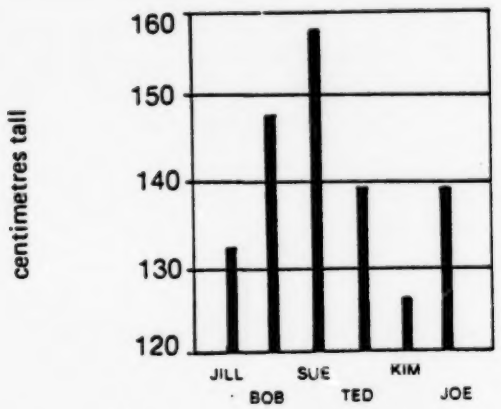
14. **Which of the following statements is correct?**

- Monday had the lowest temperature ☐ 1
- Highest temperature was recorded on Thursday ☐ 2
- It snowed on Tuesday ☐ 3
- On Wednesday the temperature was higher than on Sunday ☐ 4
- I don't know ☐ 5

15. **Some children were trying to find out which of three light bulbs was brightest. The children said the following things. Which of these gives the best start toward finding the answer?**

- One bulb looks brightest to me so I already know the answer ☐ 1
- All the bulbs look bright to me so there can't be an answer ☐ 2
- It would help if we had a way to measure the brightness of a light bulb ☐ 3
- Let's take a vote; each person will vote for the bulb he thinks is the brightest ☐ 4
- I don't know ☐ 5

16. **This is a graph of how tall 6 children are (in centimetres). Which child is shortest?**



- Jill ☐ 1
- Bob ☐ 2
- Sue ☐ 3
- Ted ☐ 4
- Kim ☐ 5
- I don't know ☐ 6

17. **Look at the graph again. Which two children are the same height?**

- Ted and Joe ☐ 1
- Sue and Ted ☐ 2
- Jill and Kim ☐ 3
- Jill and Bob ☐ 4
- I don't know ☐ 5

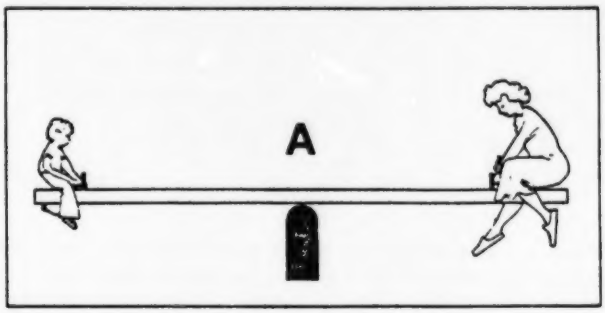
18. **If you want to find out how much a person grew in one year, which of the following MUST you know about the person?**

- His age ☐ 1
- The type of food he eats ☐ 2
- His height at the start of the year ☐ 3
- The height of his mother and father ☐ 4
- I don't know ☐ 5

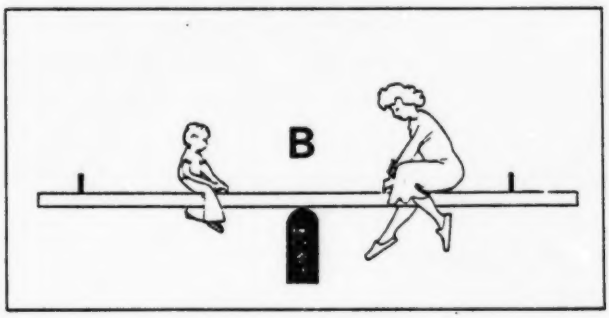
19. **Which of the following have probably been on Earth the shortest time?**

- Alligators ☐ 1
- Dragonflies ☐ 2
- Fish ☐ 3
- Men ☐ 4
- Snails ☐ 5
- I don't know ☐ 6

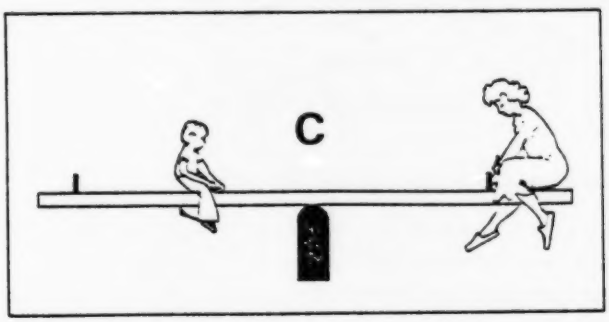
20. **Betty wanted to balance on the teeter–totter with her little brother. Which is the best way for them to sit?**



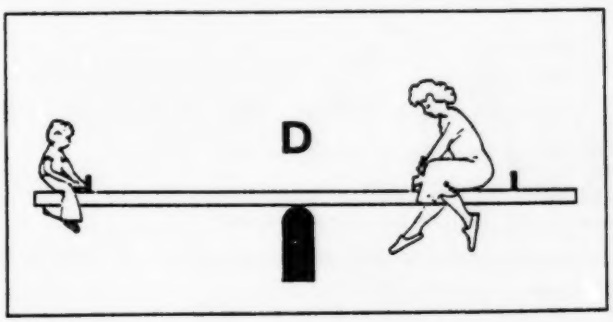
A..... ☐ 1



B..... ☐ 2



C..... ☐ 3



D..... ☐ 4

I don't know ..... ☐ 5

21. **It is most likely to snow when the weather is**

- warm and clear ☐ 1
- warm and foggy ☐ 2
- cold and clear ☐ 3
- warm and cloudy ☐ 4
- cold and cloudy ☐ 5
- I don't know ☐ 6

22. **Alice has a flat tire on her bicycle. She pumps up the tire with an air pump and begins to ride. In a few minutes, the tire is flat again. To fix her tire, Alice must find**

- a better air pump ☐ 1
- whether the tire is made of rubber ☐ 2
- where the air leaks out of the tire ☐ 3
- how many minutes it takes the tire to go flat ☐ 4
- I don't know ☐ 5

Look at the picture.

23. **Which choice tells all that you can see in the picture and no more?**

- A hand is holding a match which is burning and giving off smoke ☐ 1
- Someone has just lit a match ☐ 2
- A hand is holding a burning match ☐ 3
- Someone is about to be burned by the match he is holding ☐ 4
- I don't know ☐ 5

24. **Three teaspoons of salt were mixed into a glass of water. How could you get back all the salt?**

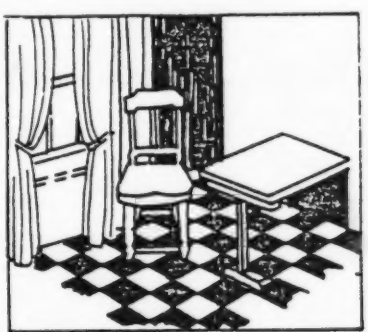
- Pour the water through a fine paper filter ☐ 1
- Use a magnet on the liquid ☐ 2
- Boil the liquid until all the water is gone ☐ 3
- Let the liquid stand overnight so that all of the salt goes to the bottom ☐ 4
- I don't know ☐ 5

A B C D

25. **Which container has a sign which means** THIS COULD POISON YOU?

A ☐ 1
B ☐ 2
C ☐ 3
D ☐ 4
I don't know ☐ 5

26. This is a picture of a room. Pretend that you are in this room and you want to measure its size, but you do not have a ruler. Which choice is **NOT** something you could use?



Your foot lengths ☐ 1
Floor tile lengths ☐ 2
Chair lengths ☐ 3
Window lengths ☐ 4
I don't know. ☐ 5

27. **Which of the following dissolves LEAST in water?**

- Glass ☐ 1
- Salt ☐ 2
- Soap ☐ 3
- Sugar ☐ 4
- I don't know ☐ 5

28. **The caterpillar is an early stage in the life history of a**

- fly ☐ 1
- bee ☐ 2
- grasshopper ☐ 3
- butterfly ☐ 4
- mosquito ☐ 5
- I don't know ☐ 6

29. **You could measure the distance from your home to the school in**

- dozens ☐ 1
- metres ☐ 2
- grams ☐ 3
- litres ☐ 4
- I don't know ☐ 5

30. **Germs sometimes make people sick because germs**

- make poisons ☐ 1
- keep the blood from moving ☐ 2
- use up all of the water in the body ☐ 3
- don't like people and want to hurt them ☐ 4
- I don't know ☐ 5

A class decided to show the temperature with the following shapes:

Circle		for over 20°C
Diamond		for 10° to 20°C
Triangle		for 5° to 10°C

31. **If the temperature at noon on Thursday is 7°C, which shape should be used?**

..... ☐ 1

..... ☐ 2

..... ☐ 3

I don't know..... ☐ 4

On each of three days the temperature at noon was:

Monday	Tuesday	Wednesday
23°C	22°C	12°C

32. **Which of the following shows the temperatures at noon for the three days?**

Mon.	Tues.	Wed.	
			..... ☐ 1
			..... ☐ 2
			..... ☐ 3
			..... ☐ 4

I don't know..... ☐ 5

A doctor kept records of breathing rates of people when they were resting. He made the chart below.

BREATHING RATES

Persons	Breaths in a minute
Baby boys	38
7-yr.-old girls	25
7-yr.-old boys	25
10-yr.-old boys	20
Mothers	16

33. This chart suggests that

- boys breathe faster than girls ☐ 1
- girls breathe faster than boys ☐ 2
- older people breathe faster than younger people ☐ 3
- younger people breathe faster than older people ☐ 4
- I don't know ☐ 5

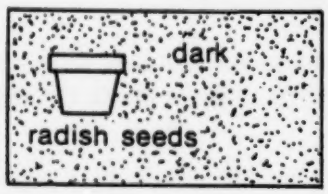
34. The bones of dinosaurs that lived millions of years ago can be found in the

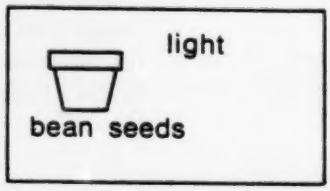
- ocean water ☐ 1
- ice on a pond ☐ 2
- trunks of trees ☐ 3
- rocks in the ground ☐ 4
- I don't know ☐ 5

35. Here is a picture of 4 objects. Which of the objects are the SAME?

- 1 and 4 ☐ 1
- 2 and 3 ☐ 2
- 1, 2, and 4 ☐ 3
- 2 and 4 ☐ 4
- I don't know ☐ 5

Tom wanted to find out whether plants can grow better in the dark or in the light. He put a pot with 6 radish seeds in a dark room and a pot with 6 bean seeds on the window sill.



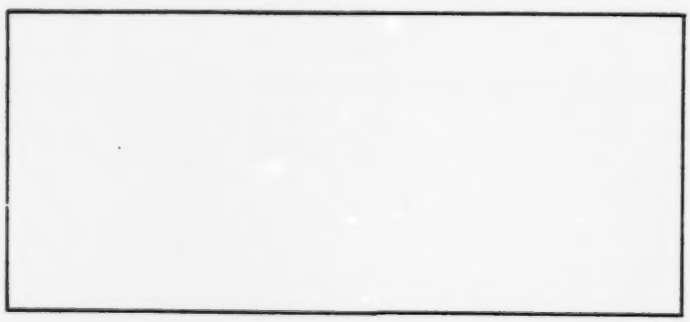


He added the same amount of water to both pots. The bean seeds grew better than the radish seeds, so Tom said his plants grow best in the light.

36. **To be able to say this, Tom should have**

watered both pots more ☐ 1
watered the radish seeds more ☐ 2
put the same kind of seeds in both pots ☐ 3
grown the seeds in water instead of soil ☐ 4
I don't know ☐ 5

37. **Use a ruler to measure the LONG side of this rectangle.**



The length is

11 centimetres ☐ 1
10 centimetres ☐ 2
9 centimetres ☐ 3
8 centimetres ☐ 4
I don't know ☐ 5

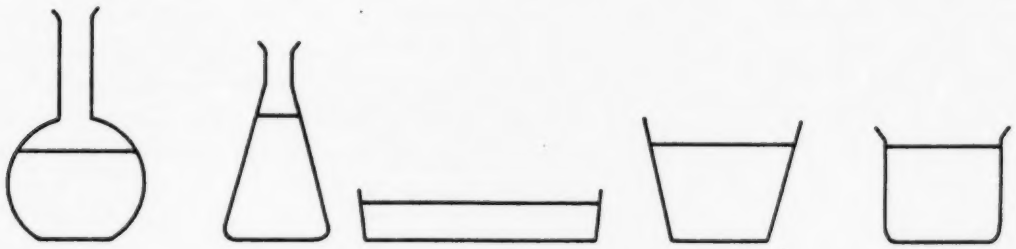
38. **Which of the following is the SUREST sign that a dog is sick?**

Its fur is clean ☐ 1
It sleeps at night ☐ 2
It is very playful ☐ 3
It has not eaten its food for two days ☐ 4
I don't know ☐ 5

39. **The reason milk kept in a refrigerator does not go sour quickly is that the cold**

changes the water of the milk into ice ☐ 1
separates the cream ☐ 2
slows down the action of bacteria ☐ 3
keeps flies away ☐ 4
causes a skin to form on the surface ☐ 5
I don't know ☐ 6

40. **A litre of water is poured into each of the containers shown below and they are left uncovered in a warm room. After a day, which container will have the LEAST amount of water left in it?**



Container A ☐ 1
Container B ☐ 2
Container C ☐ 3
Container D ☐ 4
Container E ☐ 5
I don't know ☐ 6

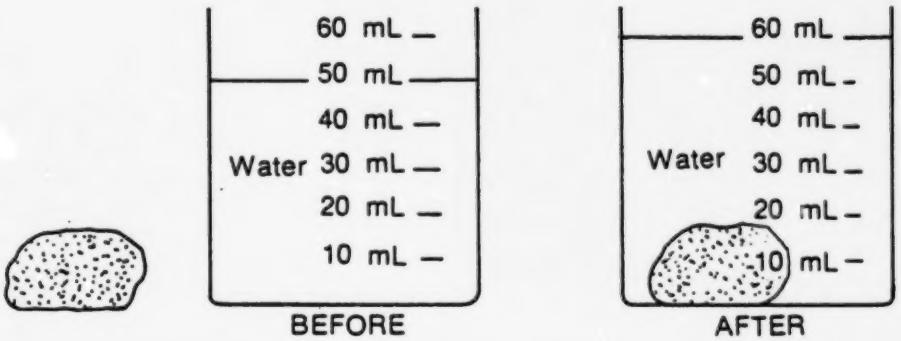
STOP

WAIT FOR INSTRUCTIONS.

SCIENCE ASSESSMENT — GRADE/YEAR 4

PART 2

A tiny rock is put into a measuring cup that has some water in it. Before the rock is put into the cup, the water is at the 50 mL line. After the rock is added, the water rises to the 60 mL line.



41. **The space taken up by the rock is**

- 10 mL ☐ 1
- 50 mL ☐ 2
- 60 mL ☐ 3
- 110 mL ☐ 4
- I don't know ☐ 5

42. **Seeds come from which one of the following parts of a plant?**

- Flower ☐ 1
- Leaf ☐ 2
- Root ☐ 3
- Stem ☐ 4
- I don't know ☐ 5

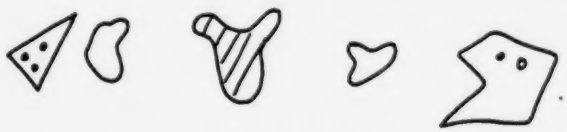
43. **If all green plants died, what would be the most important effect on man?**

- He would have to eat meat only ☐ 1
- Sooner or later man would die of starvation ☐ 2
- Man would get a little sick because he couldn't get vitamins ☐ 3
- Man couldn't build houses because there would be no lumber ☐ 4
- The land would be bare and not very pretty for man to look at ☐ 5
- I don't know ☐ 6

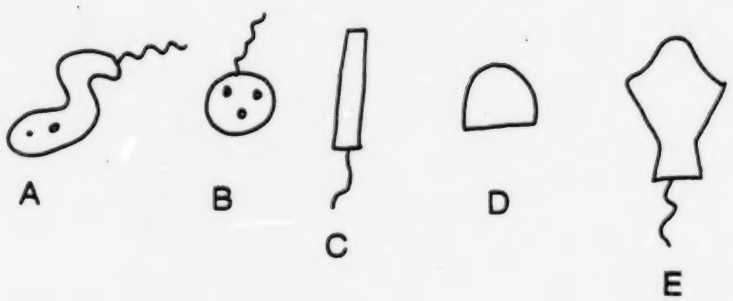
Look at the imaginary tiny animals below.

Wibbles look like those in the first row

but none of those in the second row is a Wibble.



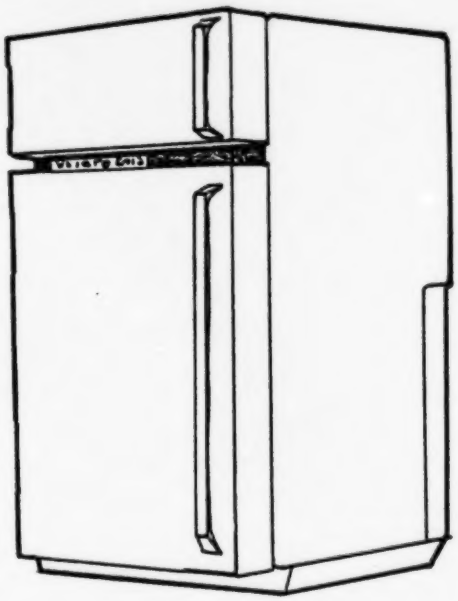
44. **Which one in the third row is NOT a Wibble?**



A..... ☐ 1
B..... ☐ 2
C..... ☐ 3
D..... ☐ 4
E..... ☐ 5
I don't know..... ☐ 6

45. **What is the MOST important thing that the lungs do?**

Hold the chest out..... ☐ 1
Protect against germs..... ☐ 2
Move different parts of the body..... ☐ 3
Pump the blood through the body..... ☐ 4
Provide a place for oxygen to enter the blood..... ☐ 5
I don't know..... ☐ 6



46. What will happen to the balloon if it is left in the refrigerator?

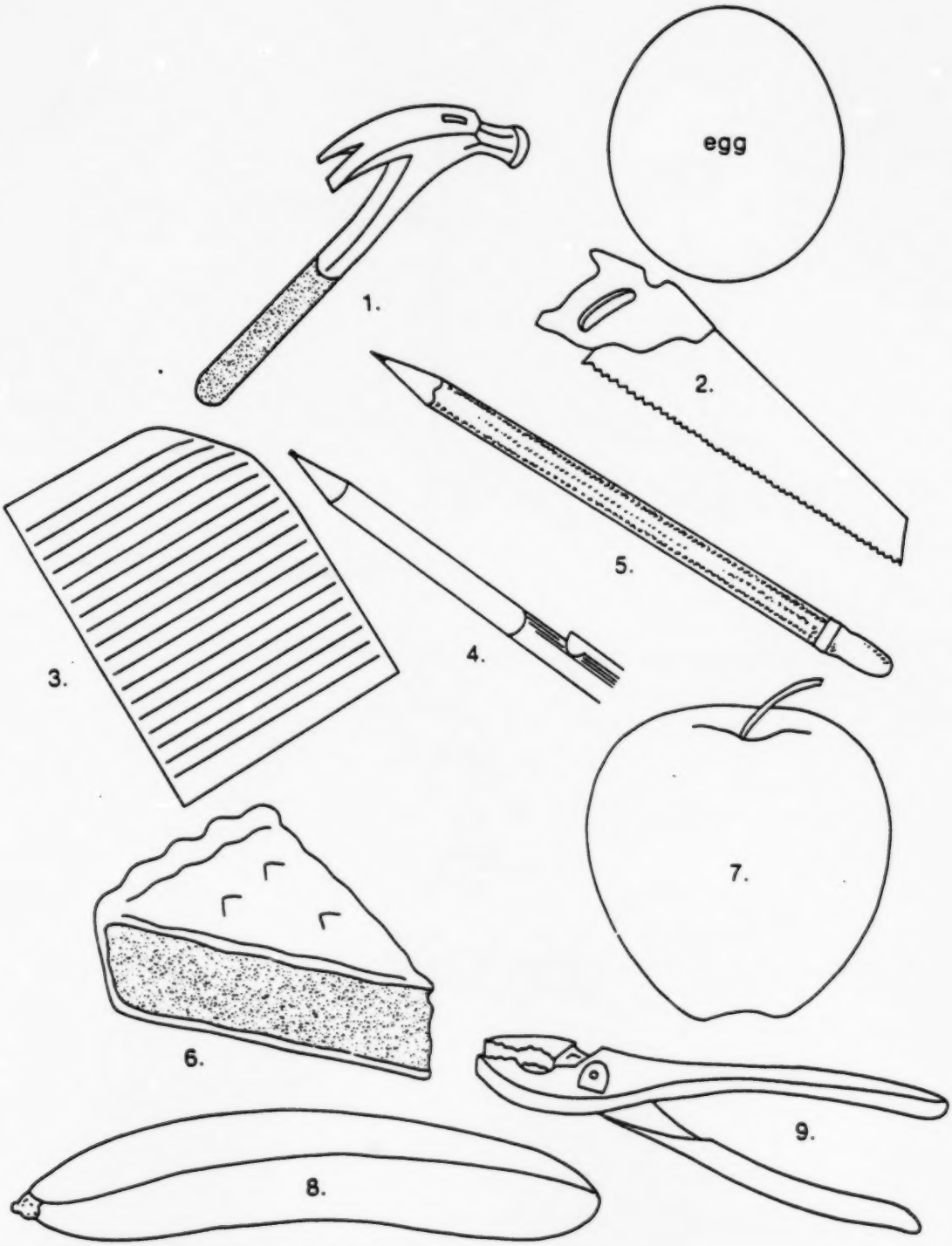
- The balloon will crack ☐ 1
- The balloon will get bigger ☐ 2
- The balloon will stay the same ☐ 3
- The balloon will get smaller ☐ 4
- I don't know ☐ 5

47. About what topic do scientists know everything?

- Clouds ☐ 1
- Stars ☐ 2
- Trees ☐ 3
- No topic ☐ 4
- I don't know ☐ 5

48. You could measure how heavy you are in

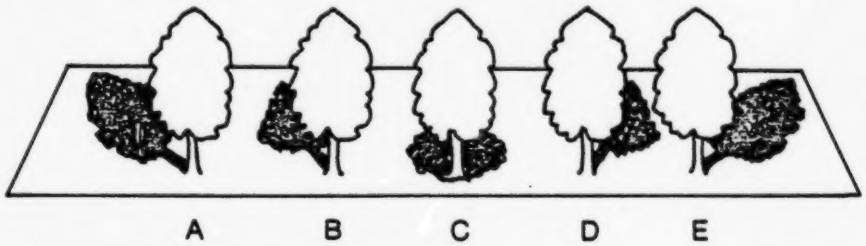
- kilograms ☐ 1
- litres ☐ 2
- kilometres ☐ 3
- millilitres ☐ 4
- I don't know ☐ 5



49 . To which set of drawings does the egg belong?

1, 2, 9	☐ 1
3, 4, 5	☐ 2
6, 7, 8	☐ 3
4, 5, 8	☐ 4
I don't know	☐ 5

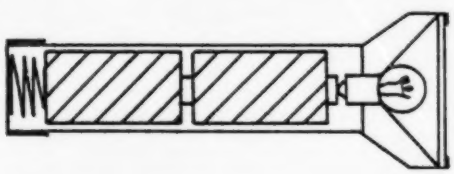
At different times during a sunny day, a tree was seen to have cast a shadow of different lengths as shown in the diagram below.



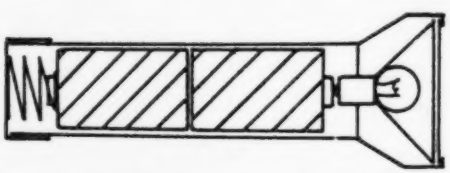
50. **Which diagram shows the shadow at midday?**

A ..... ☐ 1
B ..... ☐ 2
C ..... ☐ 3
D ..... ☐ 4
E ..... ☐ 5
I don't know ..... ☐ 6

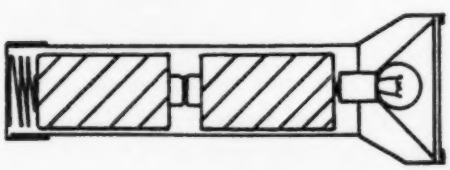
51. **Look at the diagrams below. To make the flashlight work, which way must we place the batteries?**



..... ☐ 1



..... ☐ 2



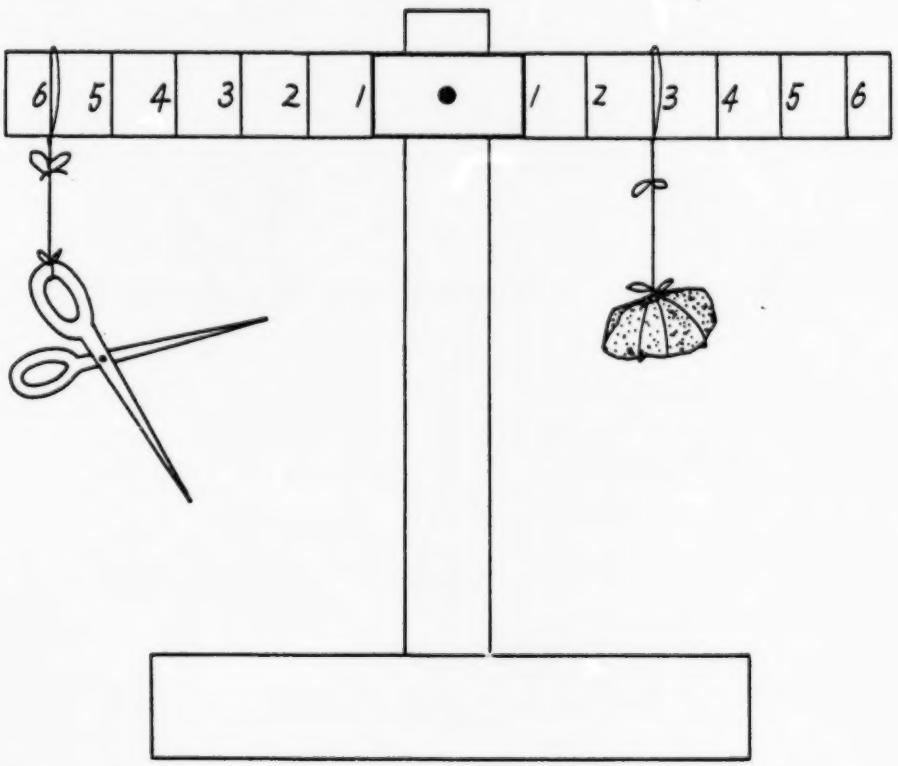
..... ☐ 3

I don't know ..... ☐ 4

52. **Iron is most likely to rust when it is**

- damp ☐ 1
- dry ☐ 2
- painted ☐ 3
- covered with soap ☐ 4
- covered with grease ☐ 5
- I don't know ☐ 6

Look at the picture carefully.



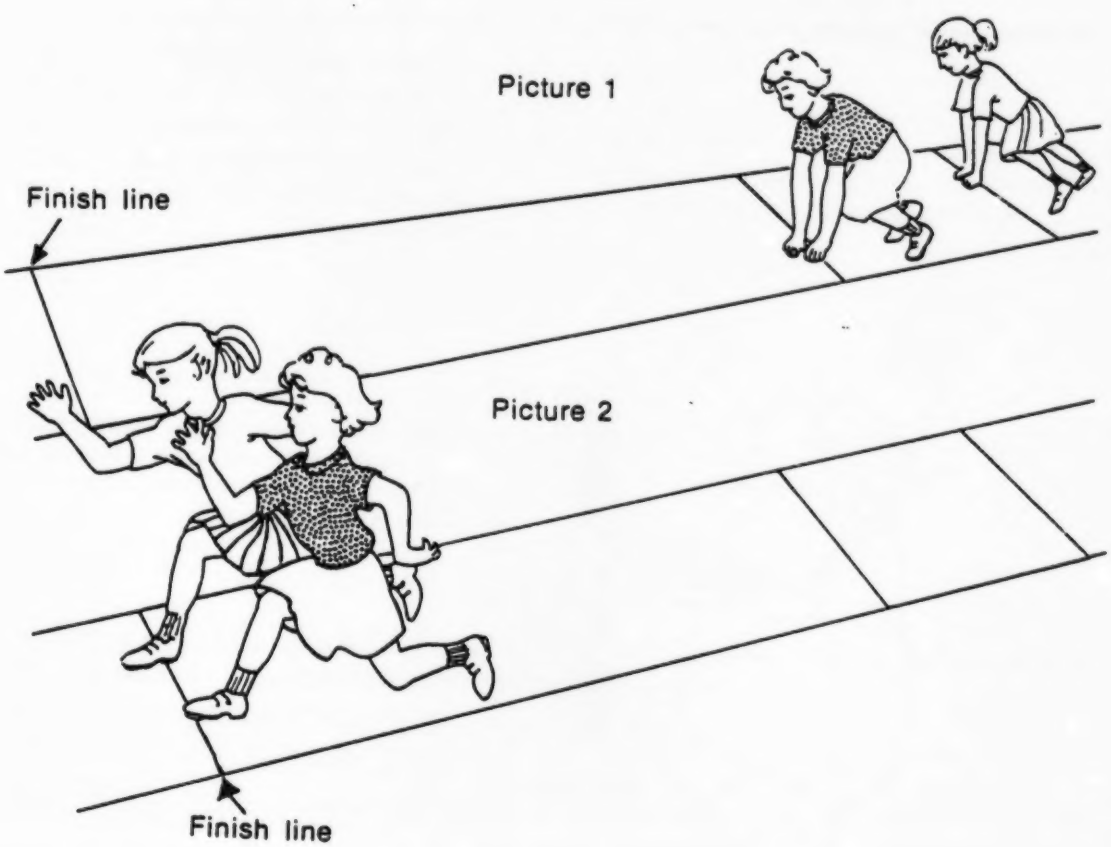
53. **From the picture you can say that**

- the scissors and the rock weigh about the same ☐ 1
- the scissors weigh about two times as much as the rock ☐ 2
- the rock weighs about two times as much as the scissors ☐ 3
- the rock weighs about three times as much as the scissors ☐ 4
- I don't know ☐ 5

54. **Which of the following COULD cause a rainbow?**

- Fog and smog ☐ 1
- Rain and snow ☐ 2
- Clouds and ice ☐ 3
- Sunshine and rain ☐ 4
- I don't know ☐ 5

Look at Pictures 1 and 2 carefully. Picture 1 is a picture of two girls at the beginning of a foot race. Picture 2 is a picture of the girls as they cross the finish line.



55. **Who ran more SLOWLY?**

- The girl in the white T-shirt ☐ 1
- The girl in the dark T-shirt ☐ 2
- They were both travelling at the same speed ☐ 3
- I don't know ☐ 4

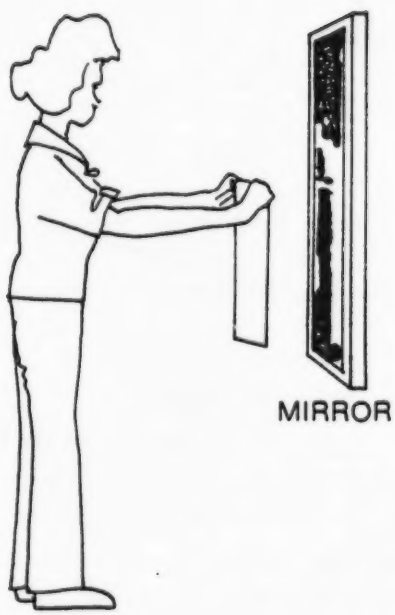
56. **Someone said that if you mix salt and sugar with water and let the mixture stand, you get saltwater taffy—a kind of candy. Which of the following would be the best way for you to test this idea?**

- Take a vote among your friends ☐ 1
- Buy some saltwater taffy and see if it has salt in it ☐ 2
- Find out if salt and sugar have the same chemicals in them ☐ 3
- Grind up some saltwater taffy to see if you can get salt, sugar, and water ☐ 4
- Try to mix salt, sugar, and water. Let them stand, and see what happens ☐ 5
- I don't know ☐ 6

57. **Scientists would have the most trouble testing which of the following?**

- I have a fever ☐ 1
- I am 2.1 metres tall ☐ 2
- My dog weighs 5 kilograms ☐ 3
- I can lift a 10 kilogram box ☐ 4
- My dog is better than your dog ☐ 5
- I don't know ☐ 6

58.



You hold a paper with the word POPSICLE facing the mirror, as shown in this picture. What does the word look like when you look into the mirror?

- ELCISPOP ☐ 1
- POPSICLE ☐ 2
- POPSICLE ☐ 3
- ELCISPOP ☐ 4
- I don't know ☐ 5

Here are two pictures of a pot of water on a stove. Picture 2 was taken 5 minutes after picture 1.

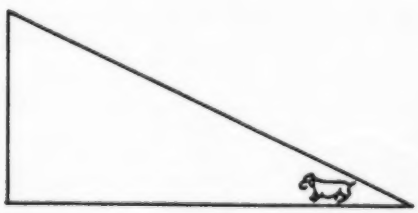
PICTURE 1

PICTURE 2

59. **Which choice is the best way of telling that there has been a CHANGE?**

The water is boiling in picture 2..... ☐ 1
The gas is on in picture 2..... ☐ 2
The water gets hot when the gas is on..... ☐ 3
The water is not boiling in picture 1..... ☐ 4
The water is boiling in picture 2, but it is not boiling in picture 1..... ☐ 5
I don't know..... ☐ 6

Look at the picture.

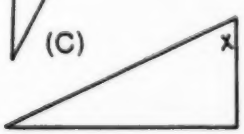


60. **It is a picture that Mary drew of a dog in a yard. Which of the following pictures shows an X in the same place in the yard as the dog in Mary's picture?**

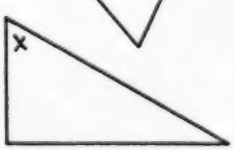
(A)

(B)

(C)



(D)



A..... ☐ 1
B..... ☐ 2
C..... ☐ 3
D..... ☐ 4
I don't know..... ☐ 5

61. **If your teacher gave you a white powder and told you to try to find out what it is, which of the following must you NEVER do (unless told to by the teacher)?**

- Put water on it ☐ 1
- Put iodine solution on it ☐ 2
- Taste it ☐ 3
- Study it under a microscope ☐ 4
- I don't know ☐ 5

62. **Mary and Dave each bought the same kind of rubber ball. Mary said, "My ball bounces better than yours." Dave replied, "I'd like to see you prove that." What should Mary do?**

- Drop both balls from the same height and see which bounces higher ☐ 1
- Throw both balls against a wall and see how far each ball bounces off the wall ☐ 2
- Drop the two balls from different heights and see which bounces higher ☐ 3
- Throw the balls down against the floor and see how high they bounce ☐ 4
- Feel the balls by hand to find which is harder ☐ 5
- I don't know ☐ 6

63. **Which one of those shown below is the best container to measure the amount of rainfall in millimetres?**

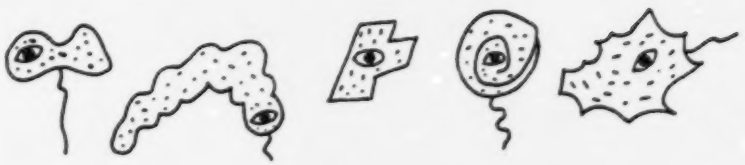
- A ☐ 1
- B ☐ 2
- C ☐ 3
- D ☐ 4
- I don't know ☐ 5

64. **How do we get scientific laws such as the law of gravity?**

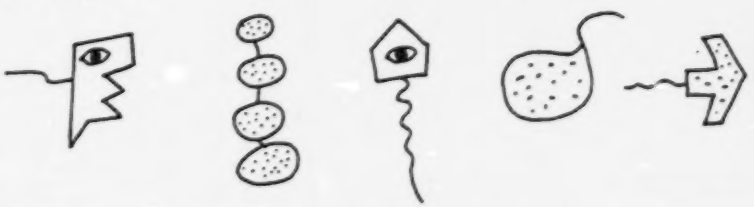
Scientists vote on them ☐ 1
They are made by the police ☐ 2
They are passed by the government ☐ 3
They come from many experiments ☐ 4
I don't know ☐ 5

Look at the pictures of imaginary pond water animals.

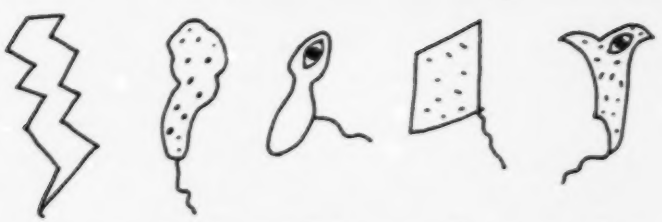
Plaps look like all those in row 1.



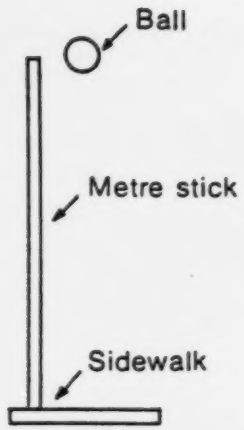
But NONE of those in row 2 is a Plap.



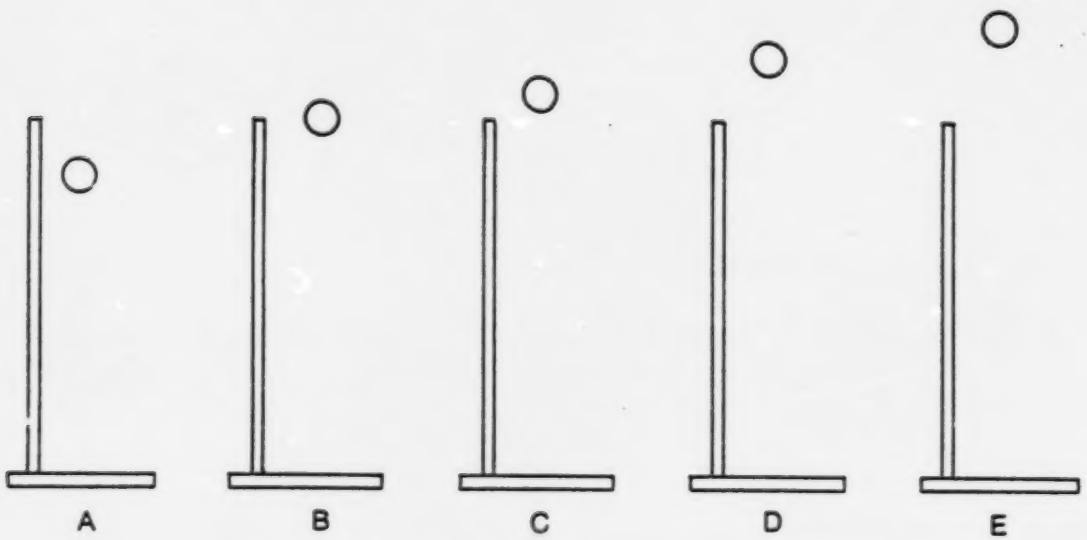
65. **Which one in row 3 is a Plap?**



A ☐ 1
B ☐ 2
C ☐ 3
D ☐ 4
E ☐ 5
I don't know ☐ 6



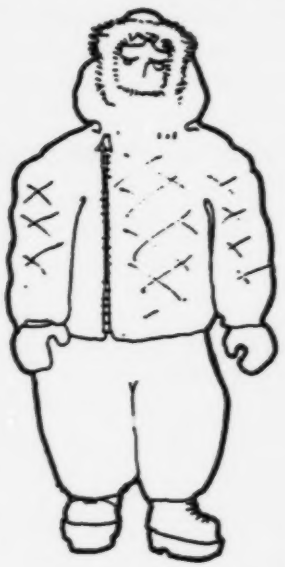
A ball is dropped on the sidewalk from a height of one metre as the picture above shows.



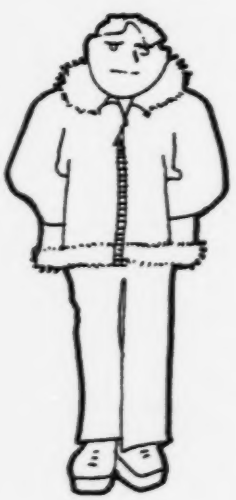
66. **Which of the above pictures shows how high the ball could bounce?**

Picture A..... ☐ 1
B..... ☐ 2
C..... ☐ 3
D..... ☐ 4
E..... ☐ 5
I don't know..... ☐ 6

67. **When the temperature outside is 25°C, how should you dress?**



A



B

C

A ..... ☐ 1
B ..... ☐ 2
C ..... ☐ 3
I don't know ..... ☐ 4

68. **After studying jars of pond water at your desk, it is important to**

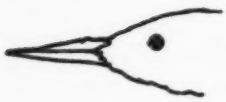
wash your hands with water..... ☐ 1
wash your hands and desk top with soap and water..... ☐ 2
spray the room with vinegar and water..... ☐ 3
stand under a hot light..... ☐ 4
I don't know..... ☐ 5

69. **While Joe was sitting under a tree, he watched a bird getting insects from between the cracks of the bark. Which drawing shows the kind of beak this bird likely had?**

A..... ☐ 1

B..... ☐ 2

C..... ☐ 3



D..... ☐ 4

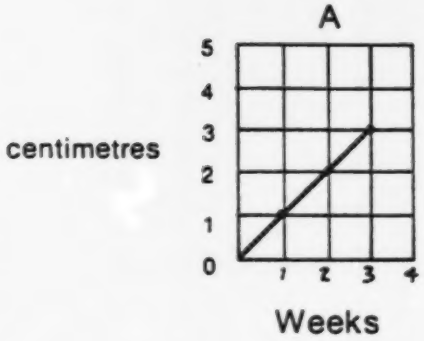
I don't know..... ☐ 5

A plant grows this way

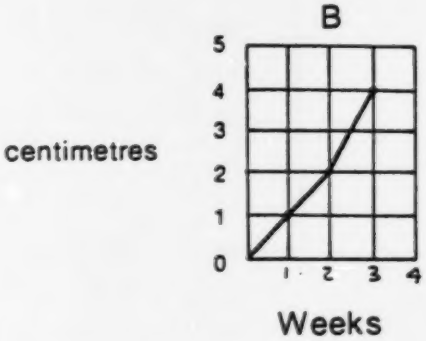
The first week it is 1 centimetre high.
The second week it is 2 centimetres high.
The third week it is 4 centimetres high.

70. **Which of the following graphs shows this growth?**

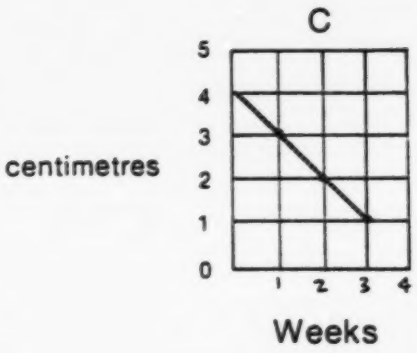
A



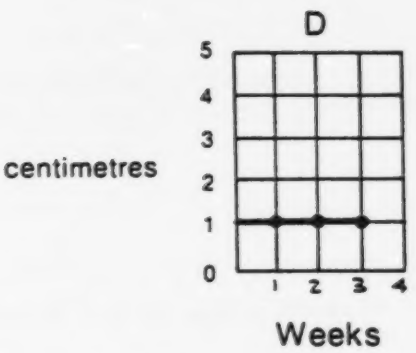
B



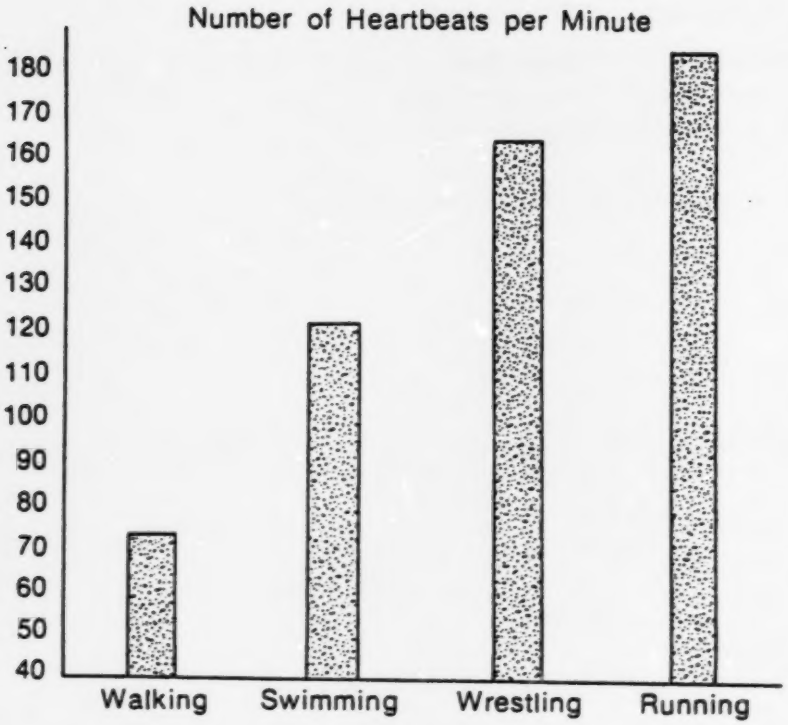
C



D



A..... ☐ 1
B..... ☐ 2
C..... ☐ 3
D..... ☐ 4
I don't know..... ☐ 5

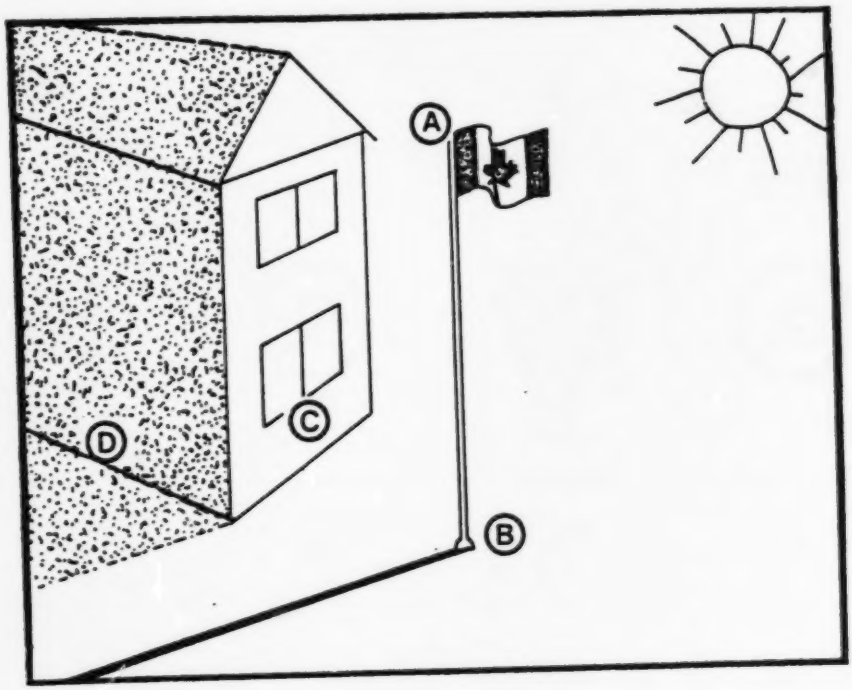


71. **Look at the graph above. Your heart probably beats fastest when you are**

walking ☐ 1
swimming ☐ 2
wrestling ☐ 3
running ☐ 4
I don't know ☐ 5

72. **When it is raining, which of the following MUST be true?**

It is spring ☐ 1
The Sun is shining ☐ 2
The wind is blowing ☐ 3
There are clouds in the sky ☐ 4
I don't know ☐ 5



73. To find the temperature of the air outdoors, it would be **best** to put the thermometer at which of the places marked in the picture?

A . . . . . ☐1
B . . . . . ☐2
C . . . . . ☐3
D . . . . . ☐4
I don't know. . . . . ☐5

74. Peter put some seeds on damp cotton in a dish. Jan put some seeds of the same kind into a glass full of water by the side of his. After two days, Peter's seeds sprouted but nothing seemed to happen to Jan's.

PETER'S SEEDS

JAN'S SEEDS

Which of the following is the most likely explanation?

Jan's seeds had been kept dry for too long . . . . . ☐1
Jan did not allow her seeds enough air . . . . . ☐2
Jan did not put the glass in a warm enough place. . ☐3
Jan should have used a different kind of seed . . . . ☐4
Jan did not use any cotton . . . . . ☐5
I don't know. . . . . ☐6

A B C D

75. **Which container has a sign which means** THIS CAN HURT YOUR SKIN AND EYES?

A ..... ☐ 1
B ..... ☐ 2
C ..... ☐ 3
D ..... ☐ 4
I don't know ..... ☐ 5

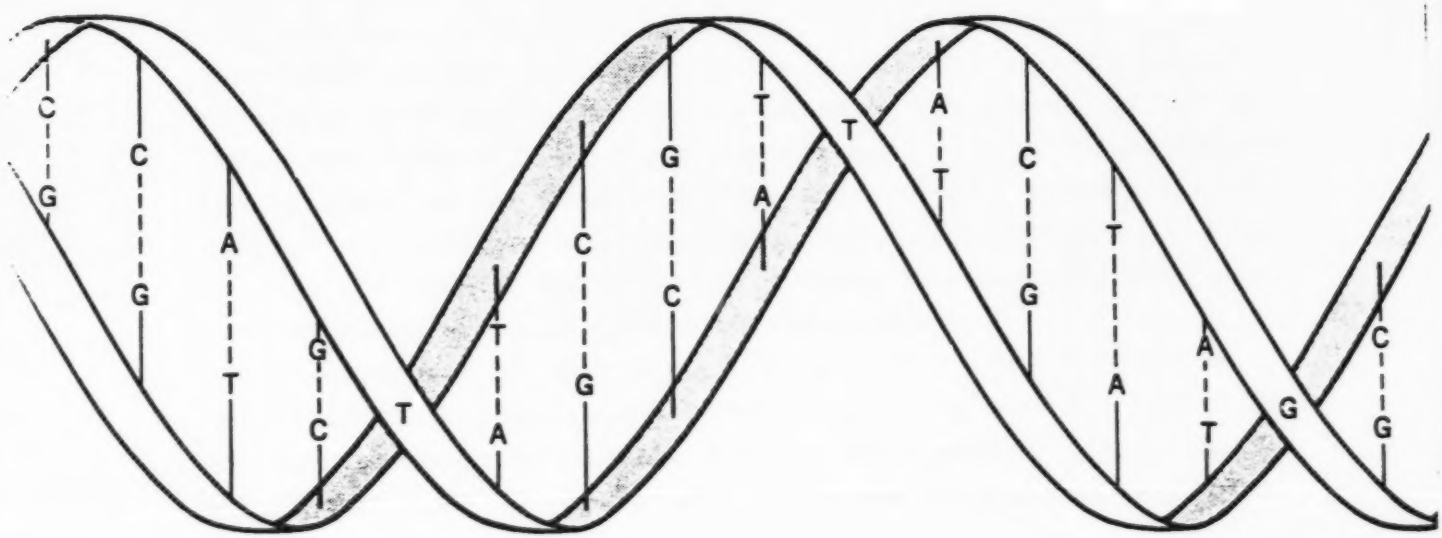
Look at these make-believe animals.

Snorps look like this:

But NONE of these is a Snorp:

76. What do Snorps have in common?

round bodies ..... ☐1
tails ..... ☐2
round bodies and tails ... ☐3
dots ..... ☐4
tails and dots ..... ☐5
I don't know ..... ☐6


C
G
C
G
A
T
G
C
T
T
A
C
G
G
C
T
A
A
T
C
G
T
A
A
T
G
C
G

APPENDIX B

CORRESPONDENCE

79-10-21

Dear Sir,

I am a Penticton intermediate teacher enrolled in a M.A. program in measurement and evaluation at U.B.C. Completion of my program requires a research project with thesis. The topic I have chosen involves the construction and comparison of tests resulting from two different approaches to item analysis - the conventional, or traditional, procedure, and the Rasch procedure. The conventional method of test construction is well researched but the Rasch model requires investigation as to its merits.

My project involves administration of a 76-item science pilot test to approximately 500 grade four students. In order to exercise greater control over the distribution and collection of test materials, I wish to base my study in the Okanagan Valley. I am requesting your permission to contact elementary school principals in your school district by letter asking their approval to administer the science test to the grade four students in their schools.

The test in question is a modified version of the pilot tests used in 1977 to construct the grade four science achievement test used in the Provincial Learning Assessment in Spring, 1978. The test will be group administered with students responding directly in the test booklets. Only one administration of the test is required with a testing time of approximately sixty minutes. I plan to have the tests ready for administration in late November or early December this year.

The purpose of the project is to compare two methods of test construction and the products resulting from the two different approaches. Individual, school, and district test data will not be computed. Consequently, strictest confidence regarding these data is assured.

I would be very grateful should you allow me to pursue this educational research in the elementary schools of your district. If your approval is granted, I would like to request a list of elementary schools in your district and their addresses. Please use the stamped, self-addressed envelope for your reply. If you

have any questions concerning this request, please call me collect at my school (492-5623). Thank you for your consideration of my request.

Sincerely yours,

John W. Knodel

John W. Knodel
215-864 Fairview Road,
Penticton, B.C.
V2A 5Y8

79-11-08

Dear Sir or Madam,

I am a Penticton intermediate teacher enrolled in a M.A. program in measurement and evaluation at U.B.C. Completion of my program requires a research project with thesis. The topic I have chosen involves construction and comparison of grade four science tests produced by applying conventional and Rasch criteria of item analysis. The nature of tests resulting from the application of the Rasch model and how they compare with conventionally constructed tests needs to be better understood. I hope that my research will contribute to this understanding.

For completion of my project, I require a sample of approximately 500 grade four students in the Okanagan Valley. I have received permission from your Superintendent of Schools to contact principals in your district asking for permission to administer a 76-item grade four science test in their schools. If you grant permission, and should your school then be included in the sample of schools which will be drawn for test administration, test booklets will be sent to your school for group administration during the first week in December.

The test to be administered is a modified version of the pilot tests used in 1977 to construct the grade four science achievement test used by the Provincial Learning Assessment Program. The test will be group administered to classes of grade four students with responses checked directly in the test booklets. Only one administration of the test is required with a testing time of approximately sixty minutes, although slower workers may need a somewhat longer time to finish. Testing can be split into two sittings, both on the same day. Tests will be computer scored.

The purpose of the research is to compare two methods of test construction, not to measure differences in the children taking the tests. Consequently, the strictest confidentiality regarding individual and school results is assured.

I will be grateful should you consent to grant me permission to include your school in the pool from which I will form my sample. Testing will take place approximately the first week

in December. Please indicate your decision regarding the role you wish your school to play in this project by completing the attached form and mailing it in the stamped, self-addressed envelope. A reply by November 16 would be appreciated. If you have any questions concerning my request, please call me collect at my school (492-5623). Thank you for your consideration of my request.

Sincerely yours,

John W. Knodel

John W. Knodel,
215-864 Fairview Road,
Penticton, B.C.
V2A 5Y8

SCIENCE PILOT TEST ADMINISTRATION - GRADE FOUR STUDENTS

- ☐ YES, grade four students in this school may serve as subjects for the administration of the science test.
- ☐ NO, grade four students in this school will not be allowed to serve as subjects for the administration of the science test.

If your answer was YES, please indicate the number of copies of the test you expect to require for administration during the first week in December.

NUMBER OF COPIES OF TEST _____.

Name of School _____

Address _____ Postal Code _____

_____ Telephone _____

Name of Principal _____

Please return this form in the stamped, self-addressed envelope.

APPENDIX C

ITEM ANALYSIS DATA

Item Analysis Data--Concepts Subtest

	Conventional		Rasch					
			Panchapakesan			Random		
Item	p-value	r_{pb}*	Mean Sq. Fit	Item Disc.	Item Diff.	Mean Sq. Fit	Item Disc.	Item Diff.
1	.55	.34	0.98	1.29	-0.03	0.99	1.22	-0.06
3	.55	.38	0.93	1.42	-0.09	0.92	1.39	-0.03
4	.45	.40	0.92	1.67	0.40	0.93	1.44	0.39
5	.30	.36	0.96	1.33	1.10	0.88	1.39	1.16
6	.88	.28	1.00	0.96	-2.25	0.89	1.25	-2.01
7	.61	.31	1.00	0.94	-0.39	0.97	1.12	-0.33
10	.35	.27	1.01	1.15	0.91	1.01	0.96	0.87
12	.29	.36	0.96	1.36	1.14	0.91	1.37	1.17
13	.43	.25	1.08	0.93	0.57	1.07	0.93	0.51
19	.39	.29	1.07	0.82	0.64	1.01	1.04	0.68
20	.46	.28	1.04	1.03	0.41	1.04	1.01	0.36
21	.85	.22	0.99	0.93	-1.76	0.98	1.01	-1.73
24	.17	.02	1.17	0.62	2.15	1.34	0.37	1.99
27	.42	.33	0.94	1.59	0.63	0.98	1.26	0.57
28	.74	.29	1.02	0.77	-1.05	1.01	0.98	-1.01
30	.34	.18	1.10	0.64	1.02	1.11	0.74	0.96
34	.79	.36	0.88	1.17	-1.50	0.84	1.33	-1.32
38	.83	.36	0.79	1.39	-1.76	0.80	1.38	-1.61
40	.55	.21	1.06	0.93	0.07	1.09	0.85	-0.03

Item Analysis Data--Concepts Subtest

	Conventional		Rasch					
			Panchapakesan			Random		
Item	p-value	r_{pb}	Mean Sq. Fit	Item Disc.	Item Diff.	Mean Sq. Fit	Item Disc.	Item Diff.
42	.48	.30	0.95	1.45	0.39	0.99	1.19	0.31
45	.60	.32	0.98	1.18	-0.25	0.99	1.12	-0.26
46	.50	.06	1.19	0.29	0.41	1.27	0.28	0.21
51	.50	.23	1.09	0.72	0.24	1.06	0.96	0.22
52	.73	.42	0.98	1.14	-1.28	0.89	1.42	-1.01
53	.19	.07	1.17	0.41	1.90	1.22	0.43	1.75
54	.86	.33	0.79	1.38	-1.98	0.79	1.26	-1.84
55	.69	.24	0.95	1.19	-0.59	1.00	0.95	-0.67
58	.36	.19	1.07	0.72	0.93	1.10	0.69	0.82
67	.31	.16	1.07	0.83	1.21	1.16	0.68	1.10
72	.76	.31	0.97	0.95	-1.22	0.96	1.06	-1.18
		Mean	1.00	1.04	0.00	1.01	1.04	0.00
		Standard Deviation	0.10	0.34	1.16	0.13	0.31	1.08

* Item/subtest point-biserial correlation coefficients corrected for spuriousness using Guilford's (1953) formula.

Item Analysis Data--Processes Subtest

	Conventional		Rasch					
			Panchapakesan			Random		
Item	p-value	r_{pb}*	Mean Sq. Fit	Item Disc.	Item Diff.	Mean Sq. Fit	Item Disc.	Item Diff.
8	.65	.40	0.96	1.12	0.10	0.95	1.09	0.07
9	.60	.24	1.18	0.71	0.41	1.23	0.64	0.31
11	.92	.27	0.93	0.69	-2.10	0.88	0.95	-2.03
14	.82	.34	0.88	1.12	-0.98	0.88	1.08	-0.99
16	.95	.29	0.81	0.95	-2.90	0.69	1.06	-2.64
17	.97	.31	0.47	1.31	-3.91	0.41	1.19	-3.45
18	.64	.37	0.98	1.04	0.13	0.96	1.07	0.12
22	.85	.44	1.39	1.21	-1.49	1.15	1.27	-1.32
23	.73	.30	1.05	0.85	-0.30	1.09	0.84	-0.38
26	.32	.32	0.97	1.14	1.81	1.11	1.02	1.73
29	.69	.29	1.09	0.87	-0.11	1.10	0.90	-0.16
31	.69	.50	0.78	1.53	-0.15	0.79	1.48	-0.14
32	.36	.32	0.98	1.16	1.59	1.03	1.10	1.53
33	.69	.45	0.90	1.13	-0.23	0.86	1.25	-0.16
35	.92	.31	0.72	1.03	-2.19	0.72	1.07	-2.16
36	.44	.37	0.96	1.13	1.12	0.99	1.13	1.08
37	.80	.32	0.93	0.93	-0.84	0.92	1.00	-0.84
41	.67	.43	0.90	1.19	-0.03	0.89	1.23	-0.04

Item Analysis Data--Processes Subtest

	Conventional		Rasch					
			Panchapakesan			Random		
Item	p-value	r_{pb}	Mean Sq. Fit	Item Disc.	Item Diff.	Mean Sq. Fit	Item Disc.	Item Diff.
44	.61	.24	1.11	0.79	0.37	1.18	0.69	0.26
48	.57	.40	0.95	1.11	0.49	0.97	1.12	0.44
49	.90	.25	1.07	0.83	-1.79	1.03	0.93	-1.79
56	.49	.45	0.92	1.22	0.84	0.91	1.29	0.84
59	.38	.34	1.02	1.04	1.42	1.05	1.00	1.38
60	.73	.30	1.08	0.73	-0.37	1.07	0.82	-0.40
62	.60	.29	1.09	0.83	0.37	1.12	0.78	0.30
63	.27	.24	1.13	0.74	2.10	1.22	0.77	2.04
65	.44	.36	0.99	1.06	1.13	1.00	1.15	1.10
70	.47	.36	1.03	1.04	1.00	1.01	1.02	0.98
71	.89	.34	0.99	1.07	-1.79	0.93	1.08	-1.74
73	.56	.09	2.06	0.10	3.85	2.53	0.13	3.69
74	.39	.41	0.95	1.21	1.37	0.91	1.26	1.35
76	.45	.43	0.95	1.24	1.07	0.93	1.20	1.04
		Mean	1.01	1.01	0.00	1.02	1.02	0.00
		Standard Deviation	0.25	0.25	1.59	0.32	0.25	1.50

* Item/subtest point-biserial correlation coefficients corrected for spuriousness using Guilford's (1953) formula.

EDUCATIONAL
RESEARCH
INSTITUTE OF
BRITISH COLUMBIA

400 515 West 10th Avenue
Vancouver British Columbia
Canada V5Z 4A8